UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Validation and analysis of the outcomes of an overlap-aware meta-analysis of genetic association statistics for T2D and related phenotypes

Máster Universitario en Bioinformática y Biología Computacional

Autor: Ruíz Rosario Mónica

Tutor: Flannick, Jason
Ponente: Díaz Uriarte, Ramón

Junio 2020

Validation and analysis of the outcomes of an overlap-aware meta-analysis of genetic association statistics for T2D and related phenotypes

Author: Mónica Ruíz Rosario
Tutor: Jason Flannick
Ponente: Ramón Díaz Uriarte

Escuela Politécnica Superior
Universidad Autónoma de Madrid
June 2020

ABSTRACT

Introduction: The amount of genetic association studies has increased exponentially in the last years. Reproducibility, however, has been challenging. False-positive associations may be, in part, the result of insufficient sample sizes. Nonetheless, the reproducibility of a real association can still be difficult to achieve due to underpowering of the replication studies, gene-environment interactions, among others. A meta-analysis as a strategy to increase the sample size, will not just find new associations but help solve the discrepancies. Nevertheless, unknown overlap may exist among the published studies, causing the inflation of the results. Therefore, this project aims to validate a new overlap aware bottom-line meta-analysis method. **Methodology:** The bottom-line analysis was executed with the meta-analysis tool Metal, and its sample overlap correction. The analysis was run over summary statistics from GWAS, WGS and WES across T2D and 58 quantitative metabolic phenotypes accessed through the type 2 diabetes knowledge portal (T2DKP). The bottom-line overlap correction was validated through the computation and comparison of its genomic inflation factor (λ) against the λ obtained by the naïve meta-analysis. The significance of the loci was confronted against the minimum p-value from the original studies, a commonly used method when accessing associations' summary statistics. **Results:** Studies with known overlap for BMI were analyzed with the bottom-line and naïve meta-analyzes. The ratio λNaïve/λBottom-line was 1.5, confirming the higher inflation on the naïve analysis results, for which the overlap was not corrected. The bottom-line discovered 26 new loci associated with BMI. The reliability of the new associations was tested by executing the analysis without the two most recently added studies to the database. As a result, 60 loci were exclusively significant for the bottom-line, from which 34 replicated in the studies not included, supporting the authenticity of the new associations. Besides being more often used the minimum p-value strategy, the ratio λMin_pvalue/λNaïve, for BMI studies with unknown overlap, was 1.7. This result suggests that even performing a naïve meta-analysis would aid on the decrement of the statistics inflation. Equivalent results were obtained for the rest of the traits. **Conclusions:** The bottom-line is a meta-analysis method capable of correcting the unknown sample overlap while improving the significance and therefore, the discovery of new real associations. A naïve meta-analysis could be an improvement against the minimum p-value to reduce the results' inflation.

RESÚMEN

Introducción: En los últimos años, se ha incrementado exponencialmente el número de estudios de asociación del genoma. Sin embargo, la reproducibilidad de los mismos ha significado un reto. Los falsos potivos reportados, podrían ser, en parte, el resultado de tamaños de muestra insuficientes. Sin embargo, la reproducibilidad de una asociación real también puede ser difícil de lograr debido a la falta de poder estadístico de los estudios de replicación, las interacciones gen-ambiente, etc. Un metaanálisis como estrategia para aumentar el tamaño de la muestra, no solo encontrará nuevas asociaciones, sino que ayudará a resolver las discrepancias. Sin embargo, puede existir una superposición desconocida entre los estudios publicados, causando la inflación de los resultados. Debido a lo anterior, este proyecto tiene como objetivo validar un nuevo método de metaanálisis basado en la superposición de resultados. **Metodología:** Se implementó un análisis "bottom-line" con la herramienta de metaanálisis Metal y su función de corrección de la superposición de muestras. El análisis se realizó sobre datos estadísticos de GWAS, WGS y WES de T2D y 58 fenotipos metabólicos cuantitativos. Los datos fueron accesados a través del portal de conocimiento de diabetes tipo 2 (T2DKP). La corrección de la superposición de las muestas se validó mediante el cálculo y comparación de su factor de inflación genómica (λ) respecto al λ obtenido por un metaanálisis "naïve" sin corrección. El valor p de los loci se enfrentó con el valor p mínimo de los estudios originales, un método comúnmente utilizado al acceder a los datos estadísticos de las asociaciones. **Resultados:** Se analizaron estudios del índice de masa corporal (IMC) con superposición de muestra conocida mediante el método bottom-line y el metaanálisis naïve. La relación λNaïve / λBottom-line fue 1.5, lo que confirma que hay una mayor inflación en los resultados obtenidos con el análisis sin corrección de la superposición. El método bottom-line permitió el descubrimiento de 26 nuevos loci asociados con el IMC. La confiabilidad de las nuevas asociaciones se analizó ejecutando el análisis sin los dos estudios agregados más recientemente a la base de datos. Como resultado, se encontraron 60 loci exclusivamente significativos para el método bottom-line, de los cuales 34 se replicaron en los estudios no incluidos, lo que respalda la autenticidad de las nuevas asociaciones. La relación λMin_pvalue / λNaïve = 1.7 sugiere que incluso realizar un metanálisis naïve ayudaría a disminuir la inflación estadística respecto la estrategia del valor p mínimo. Se obtuvieron resultados equivalentes para los demás rasgos. **Conclusiones:** el método bottom-line es capaz de corregir la superposición de muestras al tiempo que mejora el tamaño muestral y, por lo tanto, el descubrimiento de nuevas asociaciones.

ACKNOWLEDGEMENTS

Thank you to my advisor Jason Flannick and Peter Dornbos, for their guidance, teaching and support during the realization of this project.

Thank you to Furkan Büyükgöl for his contributions to the quality control of the bottom-line analysis.

Thank you to Jeffrey Massung for the implementation of the bottom-line analysis and sharing with me the methodology so that I could use it to run the analyses.

Thank you to NIMGenetics and my advisor J. C. Cigudosa for allowing me to join this project as a research internship while being part of the H2020 ITN TREATMENT (Grant Agreement No GA 721236).

INDEX

INDEX OF TABLES AND FIGURES

ABBREVIATIONS

AWS CLI: Amazon Web Services Command Line Interface

BMI: Body Mass Index

CHOL: Cholesterol

DAPPLE: Disease Association Protein-Protein Link Evaluator

GAS: Genetic Association Studies

GWAS: Genome-wide association studies

MAF: Minor Allele Frequency

NGS: Next Generation Sequencing

PPI: Protein-Protein Interaction

SNVs: Single-nucleotide variant

T2D: Type 2 Diabetes

T2DKP: Type 2 Diabetes Knowledge Portal

TG: Triglycerides

WES: Whole-Exome Sequencing

WHR: Waist-Hip Ratio

1. INTRODUCTION

Common age-related affections such as diabetes[1], hypertension[2], dementia[3], among others constitute a significant public health burden and presume to have both genetic and environmental contributions. The discovery of genetic profiles for the prediction, prevention and treatment of complex diseases has, therefore, become an increasing priority[4]. Although approximately 1,200 disease-causing genes were characterized during the 1990s decade using gene-mapping thecniques[5], the same approach has been unsuccessful for the understanding of the molecular basis of diseases that have a genetic component but do not follow Mendelian laws[4]. Genome-wide association studies (GWAS), which identify genotype-phenotype associations by testing hundreds of thousands to millions of genetic variants in many individuals[6], have today become the dominant association study design for complex diseases. As of January 2019, 3,730 GWAS had been published from which 52,415 single-nucleotide variant (SNV) – trait associations were described[6].

Even when thousands of loci have been associated with hundreds of traits and diseases' predisposition, they only explain less than half of the heritability[7]. Because GWAS mainly evaluate common variants with a minor allele frequency (MAF) > 5%, it has been hypothesized that the analysis of rare variants (MAF < 5%) could explain some of the missing heritability[8]. Since, rare variants, remain poorly characterized[9], exome chip studies were designed as a source of lower-frequency variants to complement GWAS[10]. The advances in more recently developed next-generation sequencing (NGS), have allowed the optimization of whole-exome sequencing (WES) for the regular analysis of common and rare variants[11]. Even when WES emerged as a targeted alternative to whole-genome sequencing (WGS)[12], WGS is nowadays increasing is popularity as a result of its broader coverage and lessening cost [11].

There exist limitations for genetic association studies (GAS); multiple testing is one of the major burdens[6]. By being able to interrogate million SNVs in a single study; it is necessary to reach a high level of significance, to account for the multiple tests. To preserve the false-positive rate at 5% for a million independent tests, a Bonferroni correction is applied P(study-wide type I error) = α/n ($\alpha = 0.05$, $n = 1 \times 10^6$) [13], which sets the threshold to $p < 5 \times 10^{-8}$ for associations to be considered significant. As a consequence, studies with small sample sizes, lack the power to detect many genetic variants with weak effects, leaving a big fraction of the heritability unexplained[14]. Increasing the sample size is a strategy to defeat this limitation. Since multiple

analyses have studied the same phenotypes; the meta-analyses, which combine the summary statistics from multiple studies[15], have become an effective method to accomplish bigger sample sizes and therefore to improve the power for the detection of genetic associations, even when they have modest effect sizes[16]. In addition to achieving a better statistical power, by meta-analyzing studies performed with different technologies, some of the individual limitations as the lack of common/rare variants representation and the coverage become alleviated.

Although the amount of genetic association studies has increased exponentially in the last few years, each of the them provides different information for each association, and while in some cases the outcomes replicate, in others they are not reproducible[17]. The lack of reproducibility can be attributed to the report of false positive associations as result of imprecise phenotyping, insufficient sample size, population stratification, use of control groups of unproven comparability, technical genotyping problems, lack of quality control, among others[18,19]. The reproducibility of a true association, however, can still be difficult to achieve due to underpower of the replication study, gene-environment interactions, genetic and phenotypic diverseness, etc[19]. In consequence, a meta-analysis will not just find new associations missed by the original studies but solve the discrepancies by providing a single measure of association strength that considers all the information available.

Collaborators from the Broad Institute, the University of Michigan, the University of Oxford, among others developed the Type 2 Diabetes Knowledge Portal (T2DKP)[20] to compile association studies to T2D and related phenotypes. Up to date, there are 99 datasets which include GWAS, WGS and WES with SNVs associations to 198 traits. The description of the studies can be observed in Table 1. Since each study provides different statistical information that not always replicate, the portal team made available a bottom-line meta-analysis which delivers single association statistics that combine the evidence of the original studies. The implementation was made with METAL, a genome-wide association's meta-analysis software developed in 2007 by Goncalo Abecasis, Yun Li and Cristen Willer[21]. Because unknown sample overlap may exist among the studies, the bottom-line makes use of the Metal's sample overlap correction feature, developed by Sebanti Sengupta and implemented by Daniel Taliun. This function estimates the number of common individuals among the studies based on their Z-statistics and corrects the variants' weights based on the estimated number of common individuals[22].

Table 1 Description of the datasets available in the T2DKP

(Modified from T2DKP[20])

Study	Phenotypes studied	Sample Size	Ancestry	Data type	Access
HERMES Heart Failure GWAS[23]	Heart failure	972.032	European	GWAS	Open access
DIAMANTE (European) T2D GWAS[24]	T2D, T2DadjBMI	89.813	European	GWAS	Open access
GIANT UK Biobank GWAS[25]	BMI, height	79.564	European	GWAS	Open access
CKDGen GWAS[26]	Blood urea nitrogen, chronic kidney disease, eGFR-creat	765.348	Mixed	GWAS	Open access
GIANT 2018 BMI, Height exome chip analysis[27]	BMI, height	718.734	Mixed	Exome chip	Open access
CKDGen GWAS - stratified UACR associations[28]	Urinary albumin-to-creatinine ratio	564.257	Mixed	GWAS	Open access
Joint T2D-CHD GWAS[29]	Coronary artery disease, T2D, coronary artery disease in type 2 diabetics, coronary artery disease in non-diabetics	526.043	Mixed	GWAS	Open access
UK Biobank T2D GWAS (DIAMANTE-Europeans Sept 2018)[24]	T2D	442.817	European	GWAS	Open access
AGEN and DIAMANTE T2D GWAS[30]	T2D	43.354	East Asian	GWAS	Open access
UK Biobank eBMD and fracture GWAS[31]	Estimated bone mineral density, Estimated bone mineral density	426.824	European	GWAS	Open access
GLGC exome chip analysis[32]	HDL cholesterol, LDL cholesterol, total cholesterol, triglycerides	347.532	Mixed	Exome chip	Open access
CKDGen GWAS - microalbuminuria associations[28]	Microalbuminuria	347.269	Mixed	GWAS	Open access
GIANT 2018 Body Fat Distribution exome chip analysis[33]	Waist-hip ratio adj BMI	344.369	Mixed	Exome chip	Open access
UK Biobank CardioMetabolic Consortium CHD GWAS[34]	Coronary artery disease (SOFT definition)	336.924	Mixed	GWAS	Open access
Chronic Inflammation GWAS[35]	Plasma C-reactive protein	312.468	European	GWAS	Open access
GIANT GWAS[36,37]	BMI, height,	253.288	Mixed	GWAS	Open access
GIANT Anthropometric 2015 Waist GWAS[38]	hip circumference, hip circumference adjusted for BMI, waist circumference, waist circumference adjusted	245.749	Mixed	GWAS	Open access

	for BMI, waist-hip ratio, waist-hip ratio adjusted for BMI				
GIANT 2017 smoking-adjusted GWAS[39]	BMI adjusted for smoking status, Waist circumference adjusted for BMI-smoking status, Waist-hip ratio adjusted for BMI-smoking status	241.258	Mixed	GWAS	Open access
ExTexT2D exome chip analysis[40]	T2D, T2DadjBMI	228.653	Mixed	Exome chip	Open access
COGENT-Kidney Consortium eGFR GWAS[41]	eGFR-creat (serum creatinine)	204.402	Mixed	GWAS	Open access
GIANT GWAS - stratified by physical activity[42]	BMI, waist-hip ratio, waist circumference	200.452	Mixed	GWAS	Open access
BioBank Japan GWAS[43]	Alanine transaminase, Alkaline phosphatase, Aspartate aminotransferase, Atrial fibrillation, Bilirubin, Blood urea nitrogen, BMI, Calcium Chloride, Total cholesterol, Creatine kinase, Creatinine, Diastolic blood pressure, eGFR-creat (serum creatinine), Fibrinogen, Gamma-glutamyl transferase, HbA1c, Hemoglobin, HDL cholesterol, Lactate dehydrogenase, LDL, cholesterol, Mean arterial, pressure, Menarche, Menopause, Open-angle glaucoma, Phosphorus, Plasma C-reactive protein, Potassium Pulse pressure, Random glucose, Serum albumin, Sodium, Systolic blood pressure, Triglycerides, T2D, Uric acid	191.764	East Asian	GWAS	Open access
GLGC GWAS[44,45]	Total cholesterol, LDL cholesterol, triglycerides, HDL cholesterol	188.577	Mixed	GWAS	Open access
Heart Rate GWAS[46]	Heart rate	181.171	Mixed	GWAS	Open access
DIAGRAM 1000G GWAS[47]	T2D, T2DadjBMI	159.208	European	GWAS	Open access
MAGIC Metabochip GWAS[48]	Fasting glucose, fasting insulin, two-hour glucose	13.301	European	GWAS	Open access

MAGIC HbA1c GWAS: Europeans[49]	HbA1c	123.665	European	GWAS	Open access
gnomAD exomes[50]		123.136	Mixed	WES	Open access
DIAGRAM Transethnic meta-analysis[51]	T2D	110.452	Mixed	GWAS	Open access
Global Urate Genetics Consortium GWAS[52]	Serum urate	110.347	European	GWAS	Open access
Body Fat Percentage GWAS[53]	Body fat percentage	100.716	Mixed	GWAS	Open access
BioBank Japan GWAS, males[43]	BMI, open-angle glaucoma	85.894	East Asian	GWAS	Open access
AAGILE GWAS[54]	Fasting glucose, fasting insulin adjusted for BMI	77.501	Mixed	GWAS	Open access
GoT2D exome chip analysis[55]	Fasting glucose, fasting insulin	7.567	Mixed	Exome chip	Open access
BioBank Japan GWAS, females[43]	BMI, open-angle glaucoma	7.239	East Asian	GWAS	Open access
70KforT2D GWAS[56]	T2D	70.127	Mixed	GWAS	Open access
Liver Function GWAS[57]	Alanine transaminase (ALT) levels, alkaline phosphatase (ALP) levels, aspartate aminotransferase (AST) levels, gamma-glutamyl transferase (GGT) levels	61.089	Mixed	GWAS	Open access
IAMDGC 2013 AMD GWAS[58]	Age-related macular degeneration, Neovascular age-related macular degeneration, Geographic atrophy	576	Mixed	GWAS	Open access
AMP T2D-GENES exome sequence analysis[59]	T2D	49.147	Mixed	WES	Open access
ADIPOGen GWAS[60]	Adiponectin levels	45.891	Mixed	GWAS	Open access
GoT2D WGS + replication[61]	T2D	44.414	European	GWAS	Open access
UK Biobank atrial fibrillation exome sequence analysis[62]	Atrial fibrillation	43.139	South Asian	WES	Open access
SUMMIT Diabetic Kidney Disease GWAS: subjects with T1D or T2D[63,64]	Chronic kidney disease, chronic kidney disease and diabetic, kidney disease, all diabetic kidney disease, late diabetic kidney disease, end-stage renal disease vs. no ESRD, eGFR-creat (serum creatinine), microalbuminuria	4.034	Mixed	GWAS	Open access
Leptin GWAS[65]	Leptin levels , leptin levels adjusted for BMI	32.161	European	GWAS	Open access

Primary angle closure glaucoma GWAS[66]	Primary angle closure glaucoma	26.454	Mixed	GWAS	Open access
MEDIA T2D GWAS[67]	T2D	23.827	African American	GWAS	Open access
MAGIC HbA1c GWAS: East Asians[49]	HbA1c	20.838	East Asian	GWAS	Open access
CKDGen 1000G GWAS - eGFRcys associations[68]	eGFR-cys (serum cystatin C)	20.063	Mixed	GWAS	Open access
JDRF Diabetic Nephropathy Collaborative Research Initiative GWAS[69]	Microalbuminuria, end-stage renal disease, chronic kidney disease, extreme chronic kidney disease, late diabetic kidney disease, all diabetic kidney disease	19.327	European	GWAS	Open access
FinnMetSeq exome sequence analysis[70]	2hr plasma free fatty acids, Acetate, Acetoacetate, Adiponectin, Alanine, Alcohol consumption, Beta-hydroxybutyric acid, Body fat percentage, BMI, Total cholesterol, Chylomicrons and XXL-VLDL cholesterol, Citrate Creatinine Diastolic blood pressure, Dicosahexaneoic acid, eGFR-creat (serum creatinine), Fasting glucose, Fasting insulin, Fasting plasma free fatty acids, Glutamine, Glycerol, Glycine, Glycoproteins, HDL cholesterol, HDL2 cholesterol, HDL3 cholesterol, Height, Hip circumference adj BMI, Histidine, IDL cholesterol, Concentration of IDL particles, Isoleucine, LDL cholesterol, Leucine, Linoleic acid, Omega-3 fatty acids, Omega-6 fatty acids, Phosphocholines, Phenylalanine, Plasma C-reactive protein, Pulse pressure, Pyruvate, Remnant cholesterol, Serum albumin, Serum ApoA1 Serum ApoB, Triglycerides, Total cholines, Total phosphoglycerides, Two-hour glucose, Two-hour insulin, Tyrosine, Total monounsaturated fatty acids, Total polyunsaturated fatty	19.291	European	WES	Open access

	acids, Systolic blood pressure, Total saturated fatty acids, Sphingomyelins, VLDL cholesterol, Valine, Vitamin D, Waist circumference adj BM, Weight				
VATGen GWAS[71]	Subcutaneous adipose tissue, volume, visceral adipose tissue volume, visceral adipose tissue volume adj BMI, pericardial adipose tissue volume, pericardial adipose, tissue volume adj height-weight, subcutaneous adipose tissue attenuation, visceral adipose tissue attenuation, ratio visceral:subcutaneous adipose tissue volume, ratio visceral:subcutaneous adipose tissue volume adj BMI	18.332	Mixed	GWAS	Open access
PGC GWAS[72]	Bipolar disorder, major depressive disorder, schizophrenia	16.731	Mixed	GWAS	Open access
gnomAD whole genomes[50]		15.496	Mixed	WES	Open access
Early Growth Genetics Consortium GWAS[73]	Childhood obesity	13.848	Mixed	GWAS	Open access
13K exome sequence analysis[61]	HbA1c, fasting glucose, fasting insulin, BMI, waist-hip ratio, height, waist circumference, hip circumference, triglycerides, total cholesterol, HDL cholesterol, LDL cholesterol, diastolic blood pressure, systolic blood pressure	12.954	Mixed	WES	Open access
Diabetic Cohort - Singapore Prospective Study - SEED - Living Biobank GWAS[74]	HbA1c, HbA1c adjusted for BMI, BMI, Creatinine, Diastolic blood pressure, eGFR-creat (serum creatinine), HDL cholesterol, LDL cholesterol, Systolic blood pressure	12.109	Mixed	GWAS	Pre-publication
TOPMed HbA1c Meta-analysis WGS[75]	HbA1c	10.338	Mixed	WES	Open access
Diabetic Cohort - Singapore Prospective Study - SEED GWAS[74]	T2D, T2DadjBMI	10.248	Mixed	GWAS	Pre-publication
GWAS SIGMA[76]	T2D	8.891	Hispanic	GWAS	Open access
MAGIC HbA1c GWAS: South Asians[49]	HbA1c	8.874	South Asian	GWAS	Open access

METSIM GWAS[77]	T2D, T2DadjBMI, fasting glucose, fasting glucose adjusted for BMI, fasting insulin, fasting insulin adjusted for BMI, HbA1c, HbA1c adjusted for BMI, serum creatinine, systolic blood pressure, HDL cholesterol, eGFR-creat (serum creatinine), diastolic blood pressure LDL cholesterol, BMI	8.493	European	GWAS	Open access
SIGMA exome chip analysis[76]	T2D	8.214	Hispanic	Exome chip	Open access
MAGIC HbA1c GWAS: African Americans[77]	HbA1c	7.564	African American	GWAS	Open access
Oxford BioBank Axiom GWAS[78]	BMI, Total cholesterol, HDL cholesterol, LDL cholesterol, log triglyceride level	7.193	European	GWAS	Open access
Oxford BioBank exome chip analysis[79]	Adiponectin, BMI, diastolic blood pressure, fasting glucose, fasting insulin, HDL cholesterol, height hypertension, LDL cholesterol pulse pressure, systolic blood pressure, total cholesterol, triglycerides, waist-hip ratio	7.193	Mixed	Exome chip	Open access
EXTEND GWAS	T2D, alanine transaminase (ALT) levels, alkaline phosphatase (ALP) levels, bilirubin, creatinine, diastolic blood pressure, diastolic blood pressure, eGFR-creat (serum creatinine), body fat percentage HbA1c (mmol/L), hip circumference log triglyceride level, potassium, ratio total to HDL cholesterol, triglycerides, sodium, systolic blood pressure, blood urea (mmol/L), waist-hip ratio, weight	7.159	European	GWAS	Open access

GoDarts Metabochip GWAS[80]	Adiponectin levels, BMI, cholesterol, creatinine, diastolic blood pressure, eGFR-creat (serum creatinine), fasting insulin, HbA1c, (mmol/L), HDL cholesterol height, LDL cholesterol, leptin levels, systolic blood pressure, triglycerides, T2D, waist ircumference, weight	7.119	European	GWAS	Open access
Hong Kong Diabetes Register GWAS[81]	BMI, End-stage renal disease in type 2 diabetics, Fasting insulin, Chronic kidney disease in type 2 diabetics, Coronary artery disease in type 2 diabetics,Coronary heart disease or stroke or peripheral vascular disease in type 2 diabetics, eGFR-creat (serum creatinine), HDL cholesterol, Height, HOMA-B, Insulinogenic index, LDL cholesterol, Microalbuminuria, Macroalbuminuria vs. controls, Total cholesterol, Triglycerides, Type 2 diabetes, Urinary albumin-to-creatinine ratio, Waist circumference	6.742	East Asian	GWAS	Pre-publication
SUMMIT Diabetic Kidney Disease GWAS: subjects with T1D or T2D, ESRD vs. controls[63]	End-stage renal disease vs. controls	5.655	European	GWAS	Open access
IVGTT-Based Insulin Secretion GWAS[82]	Acute insulin response, Acute insulin response adj SI, Acute insulin response adj BMI-SI, Disposition index adj BMI, Insulin secretion rate, Insulin secretion rate adj BMI, Peak insulin response, Peak insulin response adj SI, Peak insulin response adj BMI-SI	5.567	Mixed	GWAS	Open access

MAGIC GWAS[83]	area under the curve (AUC) for insulin, ratio of area under the curve (AUC) for insulin:AUC for glucose, corrected insulin response, corrected insulin response adjusted for Matsuda ISI, disposition index fasting glucose, fasting insulin, HbA1c, HOMA-B, HOMA-IR, incremental insulin at 30 min during oral glucose tolerance test, insulin at 30 min during oral glucose tolerance test, insulin at 30 min during oral glucose tolerance test adjusted for BMI, Matsuda insulin sensitivity index (ISI), modified Stumvoll insulin sensitivity index (ISI) adjusted for age and sex, modified Stumvoll insulin sensitivity index (ISI) adjusted for age, sex, and BMI, modified Stumvoll insulin sensitivity index (ISI) adjusted for genotype-BMI interaction, proinsulin levels, two-hour glucose, two-hour insulin	5.318	European	GWAS	Open access
FinnMetSeq exome sequence analysis: WHR adj BMI associations, females[70]	Waist-hip ratio adj BMI	4.927	European	WES	Open access
GoDarts exome chip analysis[80]	Adiponectin levels, BMI, cholesterol, creatinine, diastolic blood pressure, eGFR-creat (serum creatinine), fasting insulin, HbA1c (mmol/L), HDL cholesterol, height, LDL cholesterol, leptin levels, systolic blood pressure, triglycerides, T2D, waist circumference, weight	4.863	European	GWAS	Open access
Singapore Prospective Study - Living Biobank GWAS[84]	Fasting glucose, fasting glucose adjusted for BMI	3.515	East Asian	GWAS	Pre-publication
CAMP GWAS[85]	Type 2 diabetes, type 2 diabetes adjusted for BMI, fasting glucose, fasting glucose adjusted for BMI, fasting insulin, fasting insulin adjusted for BMI, HbA1c, HbA1c adjusted for BMI, diastolic	3.419	Mixed	GWAS	Open access

	blood, pressure, BMI, systolic blood pressure, HDL cholesterol, LDL cholesterol				
Hoorn DCS 2018[86]	BMI, cholesterol, creatinine, diastolic blood pressure, eGFR-creat (serum creatinine), HDL cholesterol, heart rate, LDL cholesterol, systolic blood pressure, triglycerides, urinary albumin-to-creatinine ratio, urinary creatinine	3.414	European	GWAS	Open access
FUSION exome chip analysis[85]	Type 2 diabetes, type 2 diabetes adjusted for BMI, fasting glucose, fasting glucose adjusted for BMI, fasting insulin, fasting insulin adjusted for BMI, diastolic blood pressure, urinary creatinine, LDL , cholesterol, HDL cholesterol, BMI, systolic blood pressure	3.400	European	Exome chip	Open access
GoDarts Affymetrix GWAS[80]	Adiponectin levels, BMI, cholesterol, creatinine, diastolic blood pressure, eGFR-creat (serum creatinine), HbA1c (mmol/L), HDL, Cholesterol, height, leptin levels, systolic blood pressure, triglycerides waist circumference, weight	2.917	European	GWAS	Open access
GoDarts Illumina Human MNIExpress GWAS[80]	Adiponectin levels, BMI, cholesterol, creatinine, diastolic blood pressure, eGFR-creat (serum creatinine), HbA1c (mmol/L), HDL, Cholesterol, height, leptin levels, systolic blood pressure, triglycerides, waist circumference, weight	2.902	European	GWAS	Open access
GENESIS GWAS[87]	Insulin sensitivity, insulin sensitivity adjusted for body mass index	2.765	European	GWAS	Open access
GoT2D WGS[61]	T2D	2.657	European	WES	Open access
FUSION Metabochip[85]	Type 2 diabetes, type 2 diabetes adjusted for BMI, fasting glucose, fasting glucose adjusted for BMI, fasting insulin, fasting insulin adjusted for BMI, diastolic blood pressure, urinary creatinine,	2.137	European	GWAS	Open access

	LDL, cholesterol, HDL cholesterol, BMI, systolic blood pressure				
Hoorn DCS 2019[86]	Fasting glucose, BMI, cholesterol, creatinine, diastolic blood pressure, eGFR-creat (serum creatinine), HbA1c, heart rate, height, HDL , cholesterol LDL cholesterol, systolic blood pressure, triglycerides, urinary albumin-to-creatinine ratio, urinary creatinine, weight	1.997	European	GWAS	Open access
Singapore Prospective Study Program GWAS[74]	Fasting insulin, fasting insulin adjusted for BMI	1.896	East Asian	GWAS	Pre-publication
GoDarts Illumina Infinium GWAS[80]	Adiponectin levels, BMI, cholesterol, creatinine, diastolic blood pressure, eGFR-creat (serum creatinine), fasting insulin, HbA1c (mmol/L), HDL cholesterol, height, LDL cholesterol, leptin levels, systolic blood pressure, triglycerides, T2D, waist circumference, weight	1.884	European	GWAS	Open access
FUSIN GWAS[85]	Type 2 diabetes, type 2 diabetes adjusted for BMI, fasting glucose, fasting glucose adjusted for BMI, fasting insulin, fasting insulin adjusted for BMI, diastolic blood pressure, urinary creatinine, LDL cholesterol, HDL cholesterol, BMI, systolic blood pressure	1.683	European	GWAS	Open access
1000 Genomes[88]		1.092	Mixed	WGS	Open access
GIANT-UK Biobank GWAS Meta-analysis[89]	Waist-hip ratio, Waist-hip ratio adj BMI, BMI	694.649	European	GWAS	Open access
CARDIoGRAMplusC4D-UK Biobank CAD GWAS Meta-analysis[90]	Coronary artery disease	547.261	European	GWAS	Open access
UK Biobank dietary habit GWAS[91]	Fresh fruit consumption (pieces of fresh fruit per day), Oily fish consumption (overall oily fish intake), Salt addition to food (frequency of adding salt to food), PC1 dietary pattern, PC3 dietary pattern	44.921	European	GWAS	Open access

UK Biobank 409K GWAS[92]	Diastolic blood pressure, Hypothyroidism, Red blood cell count, Systolic blood pressure	408.963	European	GWAS	Open access
UK Biobank Mendelian trait GWAS[93]	Calcium, Bilirubin, Random glucose, LDL cholesterol, Triglycerides	380.228	European	GWAS	Pre-publication
GIANT-UK Biobank GWAS Meta-analysis: females[89]	Waist-hip ratio, Waist-hip ratio adj BMI, BMI	379.501	European	GWAS	Open access
GIANT-UK Biobank GWAS Meta-analysis: males[89]	Waist-hip ratio, Waist-hip ratio adj BMI, BMI	315.284	European	GWAS	Open access
UK Biobank CAD GWAS[90]	Coronary artery disease	296.525	European	GWAS	Open access
PR interval 1000G GWAS[94]	PR interval	293.051	Mixed	GWAS	Open access
Hypertension exome chip analysis[95]	Diastolic blood pressure, Hypothyroidism, Red blood cell count, Systolic blood pressure	192.763	Mixed	Exome chip	Open access
FinnGen GWAS[96]	Atrial fibrillation, Chronic kidney disease, Diabetic nephropathy, Diabetic retinopathy, Geographic atrophy, Heart failure, Hypertension, Intracerebral hemorrhage, Ischemic stroke, Neovascular age-related macular degeneration, Neuropathy in type 2 diabetics, Obesity, Open-angle glaucoma, Type 1 diabetes, Type 2 diabetes	96.499	European	GWAS	Pre-publication
UK Biobank Cardiac MRI LV GWAS[97]	Left ventricular end-diastolic volume, Left ventricular end-diastolic volume (BSA-indexed), Left ventricular ejection fraction, Left ventricular end-systolic volume, Left ventricular end-systolic volume (BSA-indexed), Stroke volume, Stroke volume (BSA-indexed)	36.041	European	GWAS	Open access
MAGIC fasting glucose change over time GWAS[98]	Fasting glucose change over time	13.807	European	GWAS	Open access
BioMe AMP T2D GWAS[99]	Type 2 diabetes, type 2 diabetes adjusted for BMI, fasting glucose, fasting glucose adjusted for BMI, HbA1c, HbA1c adjusted for BMI, serum creatinine, systolic blood pressure, HDL	9.361	Mixed	GWAS	Open access

	cholesterol, eGFR-creat (serum creatinine), diastolic blood pressure, LDL cholesterol, BMI				
Diabetic retinopathy GWAS: Europeans[100]	Diabetic retinopathy, Proliferative diabetic retinopathy, Non-proliferative diabetic retinopathy, Proliferative diabetic retinopathy	3.246	European	GWAS	Open access
Diabetic retinopathy GWAS: African Americans[100]	Diabetic retinopathy, Proliferative diabetic retinopathy, Non-proliferative diabetic retinopathy, Proliferative diabetic retinopathy	2.611	African American	GWAS	Open access

2. OBJECTIVES

2.1. GENERAL OBJECTIVE

The present project aims to perform quality control of the bottom-line analysis. Particularly, evaluate the method's capacity to account for the sample overlap and the reliability of the association outcomes.

2.2. PARTICULAR OBJECTIVES

1. Evaluate and compare the genomic inflation obtained by the results provided by the bottom-line analysis, the naïve meta-analysis and the minimum p-value method.
2. Evaluate the effect of the bottom-line analysis on the significance of the variants.
3. Evaluate the reliability of the associations established by the bottom-line analysis.
4. Analyze the biological meaning of the gained and lost associations.

3. METHODS

3.1. DATA

Summary statistics from 94 GWAS, WGS and WES datasets across T2D and 57 quantitative metabolic phenotypes were accessed through the type 2 diabetes knowledge portal (T2DKP)[20]. All data is stored in the cloud-based Amazon Simple Storage Service (Amazon S3). Data was downloaded via the Amazon Web Services Command Line Interface (AWS CLI) and all analyses were completed using the Broad Institute cluster computational resources. In all cases, traits selected had at least 2 independent studies available. Range of datasets per trait 2-25.

3.2. META-ANALYSES IMPLEMENTATION

3.2.1. BOTTOM-LINE META-ANALYSIS IMPLEMENTATION

The bottom-line analysis was implemented, as shown in Figure 1. For a trait, the variants derived from the available studies were filtered by ancestry and MAF. Multi-allelic variants and variants with missing p-value or beta/OR were removed from the analysis. If more than one ancestry were present, the "mixed" ancestry was removed from any further analyses. The common variants (MAF < 5%) from each ancestry were meta-analyzed with METAL, using the sample overlap correction and the sample size weighted scheme. A second METAL analysis without overlap correction was performed with the standard error scheme, and the output was combined with the sample size results. The rare variants from each ancestry were united with their corresponding common variants' METAL output. If any variant existed as both common and rare, only the one with the largest total N was kept. A final METAL analysis without overlap correction was performed with the variants from the different ancestries. The output p-values were considered as the bottom-line result. The code for the implementation is shown in the supplemental material 7.2. Bottom-line meta-analysis implementation code.

3.2.2. NAÏVE META-ANALYSIS IMPLEMENTATION

For a trait, a METAL analysis without overlap correction was performed with the available studies. In all cases, the selection of the studies followed the same rules as the bottom-line implementation, in terms of ancestries and subsets included in the analysis.

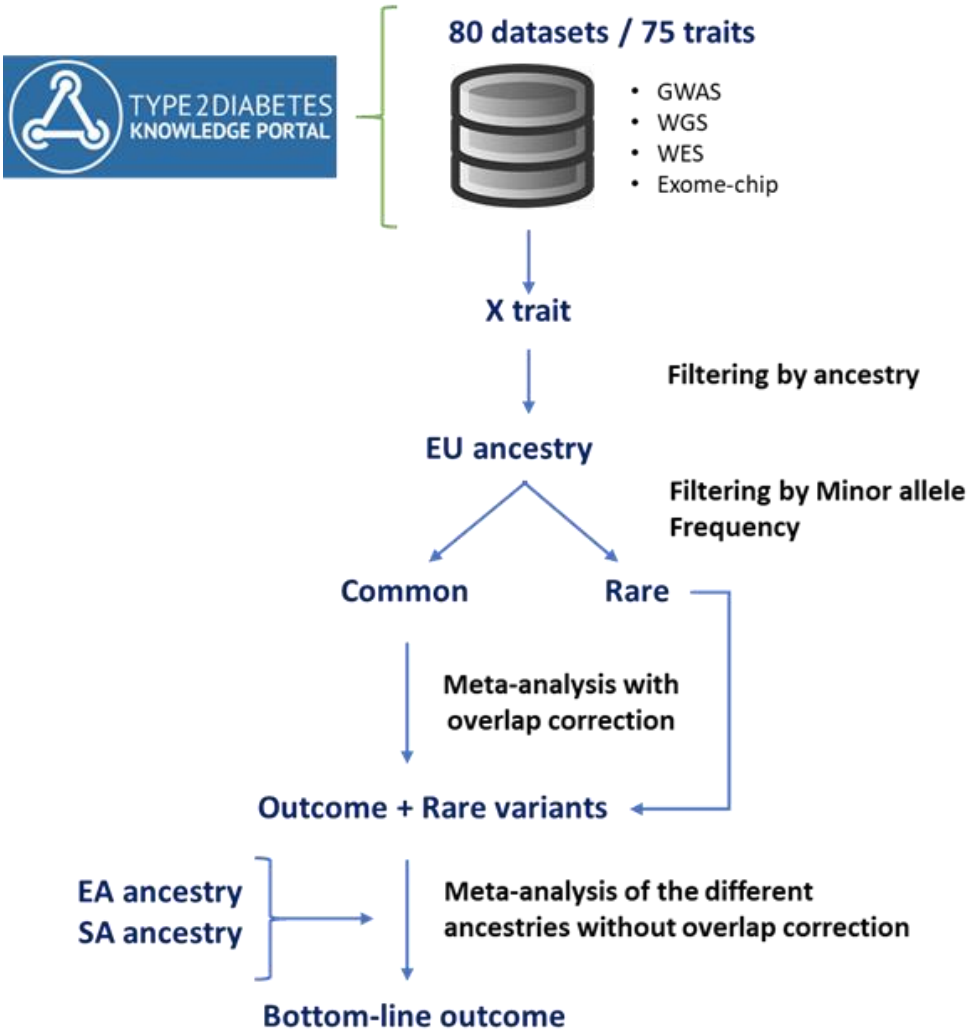


Figure 1. Bottom-line analysis implementation.
Summary statistics from GWAS, WGS and WES across T2D and other related phenotypes were accessed from the T2DKP. For each trait, the variants were filtered by MAF and ancestry. If more than one ancestry were present, the "mixed" ancestry was removed from the analysis. A METAL analysis with overlap correction was run for the common variants of each of the ancestries. The output was united with the rare variants and a final METAL analysis was run without overlap correction.

3.2.3. MINIMUM P-VALUE ANALYSIS

For each variant, the minimum p-value across all studies available for a specific trait was computed. In all cases, the selection of the studies followed the same rules as the bottom-line implementation, in terms of ancestries and subsets included in the analysis.

3.3. SAMPLE OVERLAP ASSESSMENT

The inflation of the meta-analyses results was assessed by the computation of the genomic inflation factor (λ) as the median of the chi-squared statistic divided by the median of the chi-squared distribution[101]. Q-Q plots were also drawn to visualize the distribution of the data in comparison with the theoretical distribution.

3.4. PROTEIN-PROTEIN INTERACTION ANALYSES

The Disease Association Protein-Protein Link Evaluator (DAPPLE) was used to annotate loci to the proteins encoded by them. Physical connectivity among proteins associated to disease was assessed and the resulting networks obtained. The annotated proteins were submitted to STRING database to perform protein-protein interaction enrichment analyses from which networks and enrichment p-values were obtained.

4. RESULTS AND DISCUSSION

4.1. QUALITY CONTROL OF THE BOTTOM-LINE SAMPLE OVERLAP CORRECTION

Unknown sample overlap may exist among the studies available for each trait. The bottom-line analysis' capacity to account for it can be assessed by comparing the inflation of the resulting p-values against those provided by a naïve meta-analysis. The genomic inflation factor (λ), which can be described as the median of the chi-squared statistic divided by the median of the chi-squared distribution[101], inform us about the number of significant p-values in comparison with the expected one. An association of half of the variants would compute a λ=1. Since it is not expected that more than half of the variants are associated with a specific trait, a λ > 1 would mean that the results are inflated, and more associations than the existing ones are being pointed out. Because a naïve meta-analysis does not correct for the sample overlap, it is anticipated that the λ computed for its results will be higher than the λ computed for the bottom-line p-values if overlap exists among the studies.

As a proof of concept, the bottom-line and naïve meta-analyses were run for the trait body mass index (BMI). The overlap among the individual studies is not known. However, it is known that specific characteristics as the gender have stratified some of the studies and that these stratified datasets are subsets of the general ones, implying that overlap exists among them. Figure 2A shows the totality of studies available for BMI and how some of them are subsets of others. The bottom-line was run for all studies available in the T2DKP at the date of October 2019. The subsets were also included, as a source of known overlap to validate if the bottom-line can account for it. Because more than one ancestry exists, the mixed ancestry was removed (as stated in the bottom-line implementation). The final studies included in the analysis are shown in Figure 2B. For a fair comparison, the naïve meta-analysis was run with the same studies as the bottom-line.

Figure 2. Available studies for BMI and studies included in the bottom-line sample overlap correction validation.

The studies available have been stratified by different characteristics as the gender, which means that some studies are in fact, subsets of the general ones (A). To assess the capacity of the bottom-line to account for the sample overlap, the main studies and their subsets were included in the analysis. Because more than two ancestries were present, the studies from "mixed" ancestry were not included (B).

Lambdas and Q-Q plots were computed for the bottom-line and the naïve meta-analyses results (Figure 3). The p-values from the naïve analysis, without overlap correction, calculated a λ=1.930, implying as expected that there is sample overlap among the studies. The corresponding Q-Q plot confirmed the inflation, showing that most of the p-values do not follow a theoretical distribution. The bottom-line analysis is supposed to correct the already known sample overlap. A λ=1.279 confirms that the correction is being made, reducing the inflation of the data significantly. The corresponding Q-Q plot shows an increment of the correlation of the points to the theoretical distribution in comparison with the ones obtained by the naïve analysis.

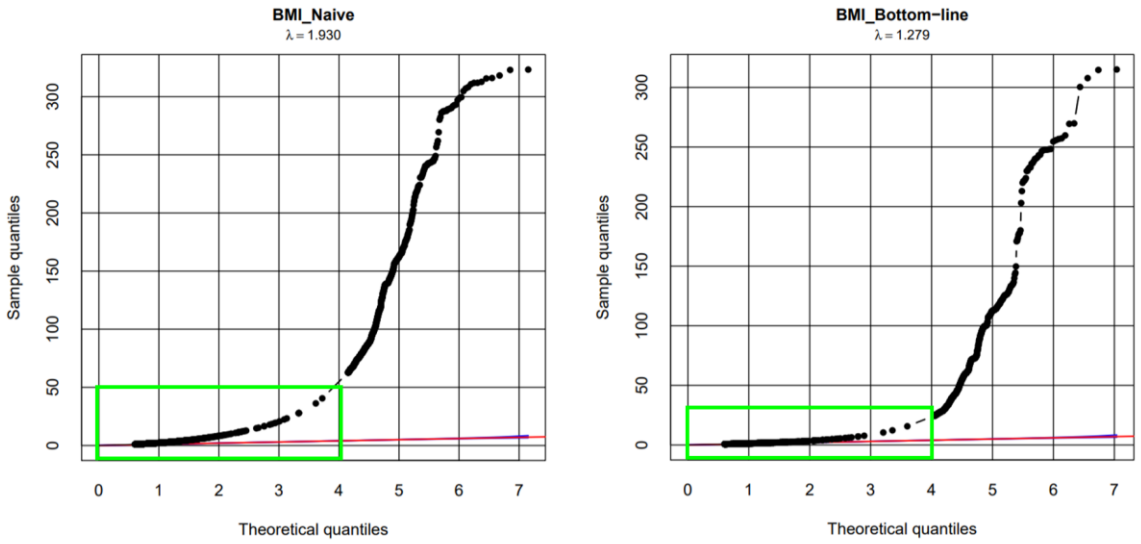


Figure 3. Genomic inflation factor (λ) and QQ-plots computed for the bottom-line and the naïve meta-analyses performed for the studies and sub-studies available for BMI.
The bottom-line and naïve meta-analyses were run over the studies and sub-studies available for BMI as a source of know sample overlap. The λ obtained for the naïve analysis was considerably higher (λ=1.930) than the one computed for the bottom-line results (λ=1.279). This outcome confirms that overlap exists among the studies meta-analyzed and that the bottom-line can account for it, reducing significantly the inflation of the results in comparison with the naïve analysis that does not correct the overlap. The QQ-plots confirm the reduction of inflation by the bottom-line. As seen inside the green rectangles, the points from the bottom-line correlate better with the theoretical distribution than the points obtained by the naïve analysis.

The bottom-line and naïve meta-analysis λ comparison was performed for the rest of the traits (Figure 4). It was observed that while for some traits the λ remained very similar for both analyses, meaning that there was not sample overlap among the studies. For other traits, the λ obtained when the overlap was not corrected (naïve analysis) was significantly higher than the one computed for the bottom-line results. These outcomes support what was observed for BMI, validating that when sample overlap exists among the studies meta-analyzed, the bottom-line analysis is being able to account for it, providing uninflated statistic results.

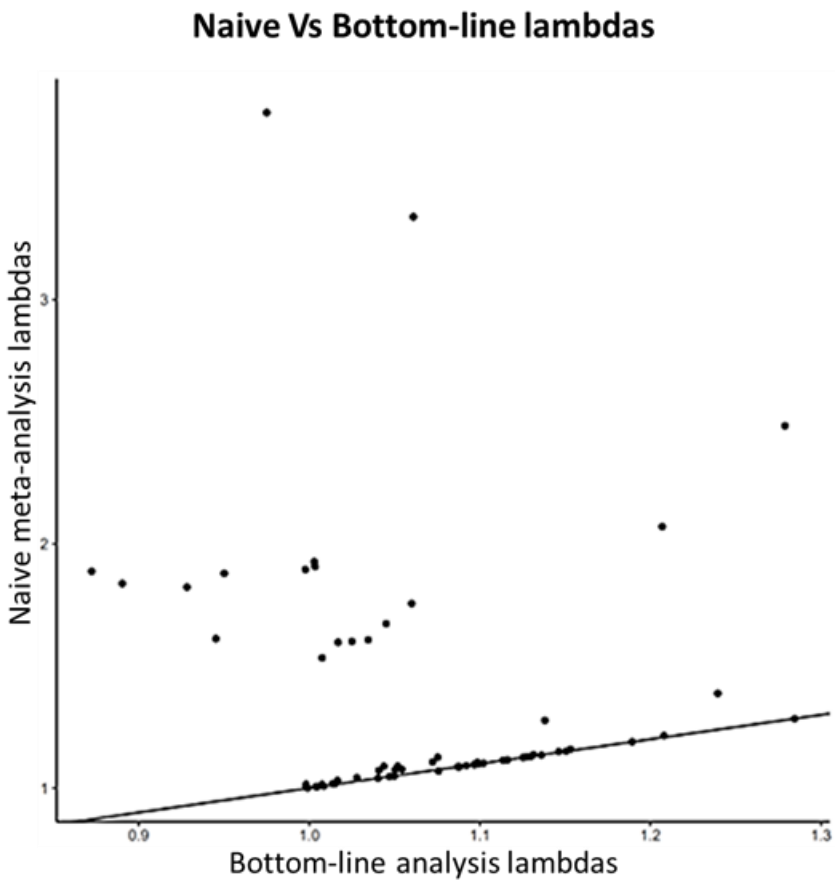


Figure 4. Comparison of the lambdas computed for the p-values obtained by the naïve and bottom-line meta-analyses when overlapping studies are present.
For several traits, the lambda obtained by the naïve meta-analysis was significantly higher than the one obtained with the bottom-line analysis. The result confirms that when overlap exists among the analyzed studies, the bottom-line can account for it, reducing the results inflation significantly.

Being already validated that the bottom-line analysis can account for the sample overlap, the inclusion of any stratified subsets was avoided in all further analyses.

Catalogues of association statistics and researchers often consider a variant to be associated if it achieves statistical significance in any published study - implicitly assigning a variant its "minimum p-value" across studies. Such an approach introduces false-positive associations into public resources and the literature. The genomic inflation factors that resulted from this approach ranged between 1.0 and 3.7 for the complex traits analyzed. The bottom-line aims to improve the reliability of the associations reported by considering all available information, instead of just the minimum p-value, while accounting for the unknown sample overlap. The bottom-line analysis corrected the genomic inflation, therefore producing well-calibrated association statistics (Figure 5A).

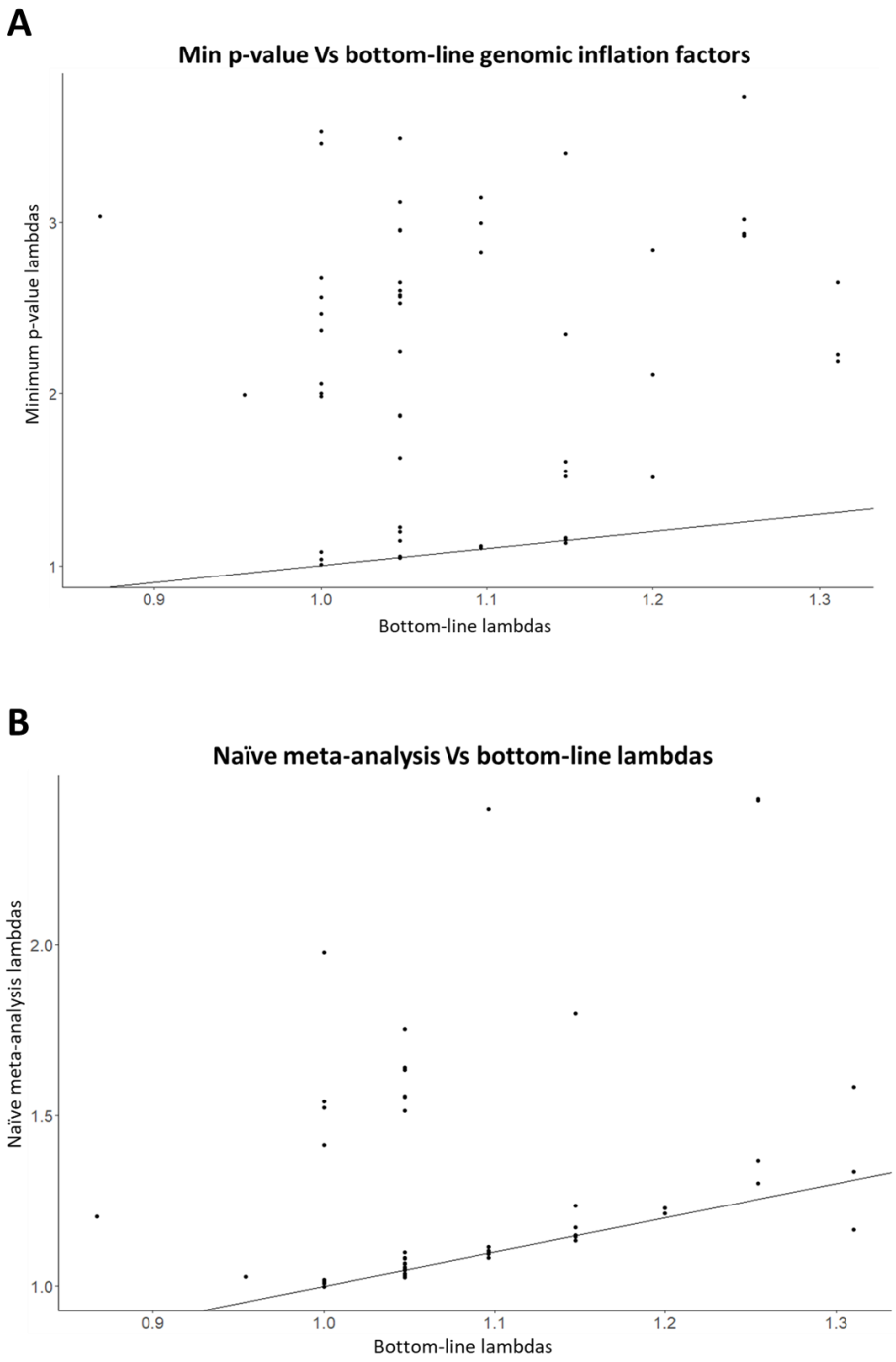


Figure 5. Comparison of the genomic inflation factors obtained by the minimum p-values, the naïve meta-analysis and the bottom-line analysis.

The genomic inflation factors for the minimum p-value analyses, which ranged between 1.0 and 3.7, became significantly corrected by the bottom-line, for which the maximum λ was 1.3. Surprisingly for 40/58 traits, the naïve meta-analysis λ was <1.1 times that of the bottom-line analysis. However, for 16 traits the λ became significantly corrected, being the maximum ratio λ naïve / λ bottom-line = 2.2.

Perhaps surprisingly, the naïve meta-analysis is reasonably well-calibrated when the fully overlapping subsets are not included. 42 traits obtained a λ<1.1 times that of the bottom-line analysis. However, correcting for sample overlap reduced the genomic inflation for 16 (28%) of

the traits, in some cases significantly (e.g. reduction from 1.8 to 1.0 for Coronary Artery Disease) (Figure 5B).

4.2. EFFECT OF THE BOTTOM-LINE ANALYSIS ON THE SIGNIFICANCE OF THE VARIANTS

17 individual studies for BMI were meta-analyzed with the bottom-line method. For each variant, the results were compared against the minimum p-values across the original studies (Figure 6). It was observed that several variants did not modify their significance. However, some variants with no significant p-values in the original studies became significant with the bottom-line analysis and, by the contrary, some variants with already reported significant associations, lost their significance after being meta-analyzed.

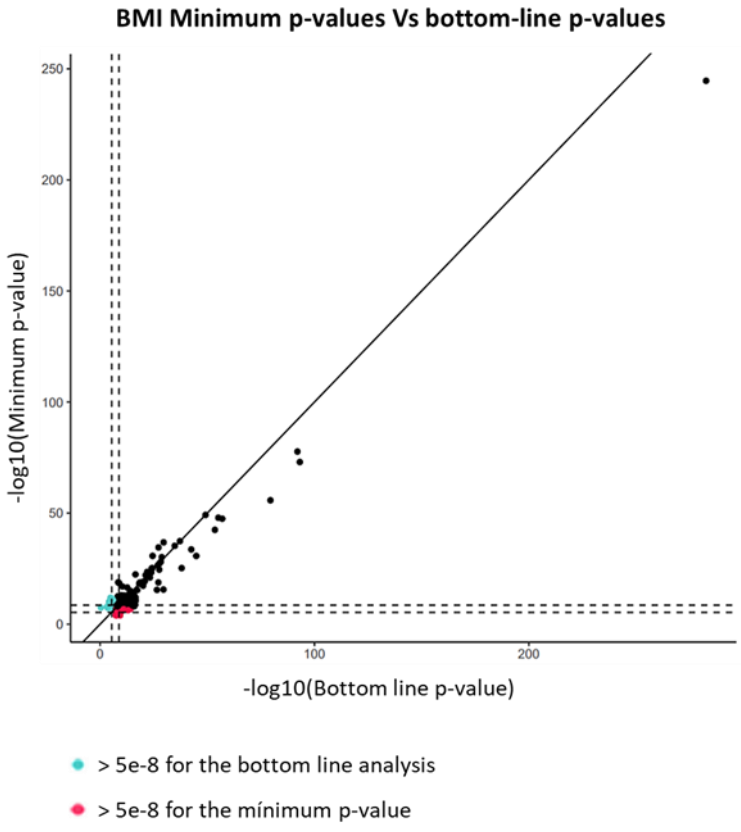


Figure 6. BMI bottom-line p-values comparison against the minimum p-values across studies
For each variant, the minimum p-values across the studies available for BMI were compared against the p-values obtained by the bottom-line analysis. Several variants did not modify their significance (black dots). However, some variants that were significant for at least one of the original studies (minimum p-value < 5e-8), lost their significance with the bottom-line analysis (blue dots). Moreover, some variants that were not initially significant (minimum p-value > 5e-8) became significant after the meta-analysis (red dots).

The number of genomic regions affected by the gain and loss of significance was assessed through the definition of loci as 150kb windows to each side of the most significant variants (lead variants). New loci were described as those containing variants exclusively significant for the bottom-line, meaning that the bottom-line is the first analysis to point their association with a trait (Figure 7A). Accordingly, lost loci were considered those with a significant p-value for at least 1 of the original studies that lost their significance when meta-analyzed with bottom-line (Figure 7B).

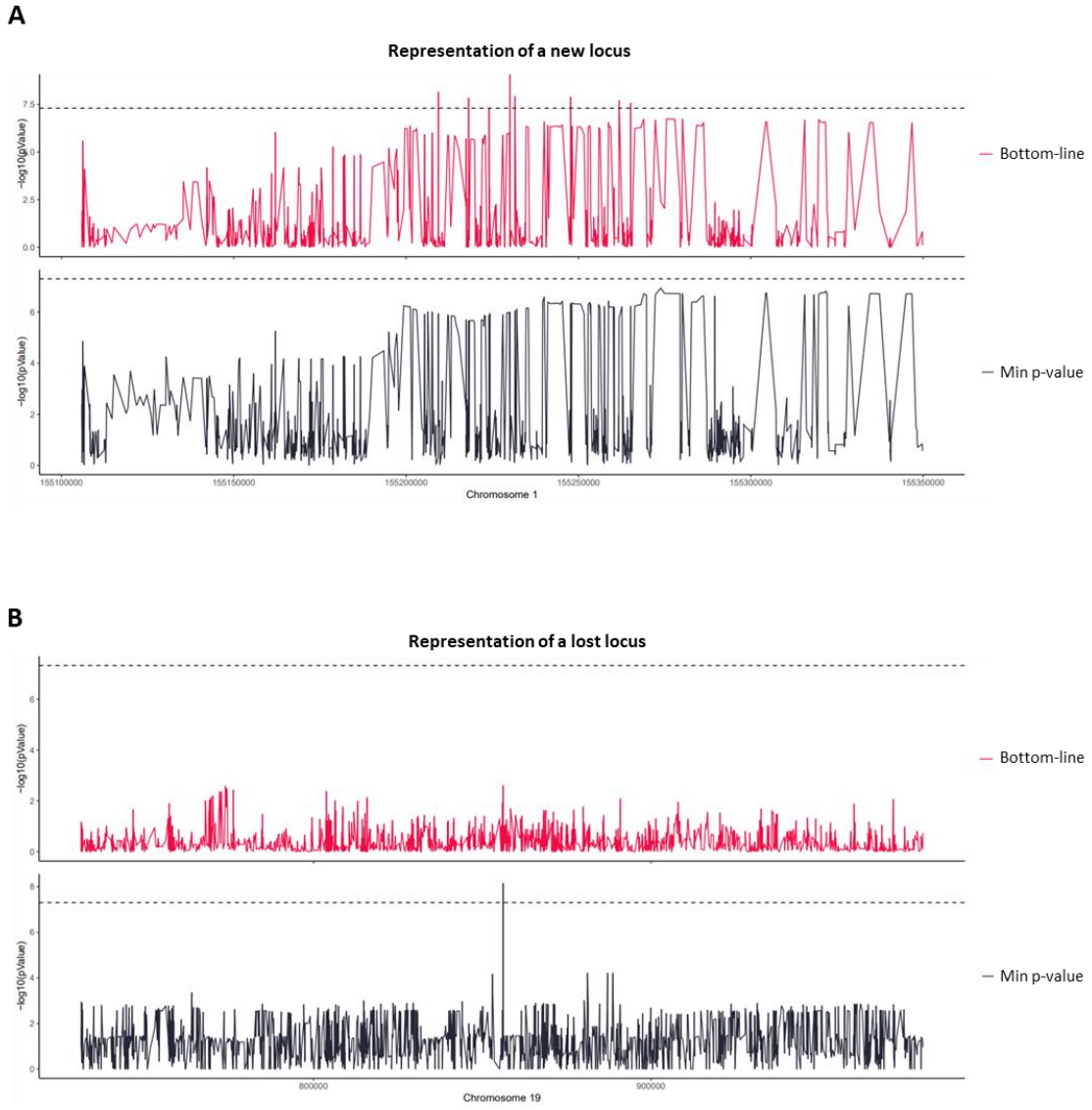


Figure 7. Graphical representation of the new and lost loci associated with a trait

The figure exemplifies the locus view of the new and lost loci associated with a trait. In a new locus, one or more of the variants have significant p-values for the bottom-line analysis, while none of them has a significant association in any of the original studies (A). In contrast, a lost locus contains variants with significant associations in at least one of the original studies, which are not significant for the bottom-line analysis (B).

The bottom-line identified 255 loci with a significant association with BMI, from which 60 were newly associated. Notwithstanding 42 loci with previously reported significant p-values, showed not significant for the bottom-line analysis (Figure 8).

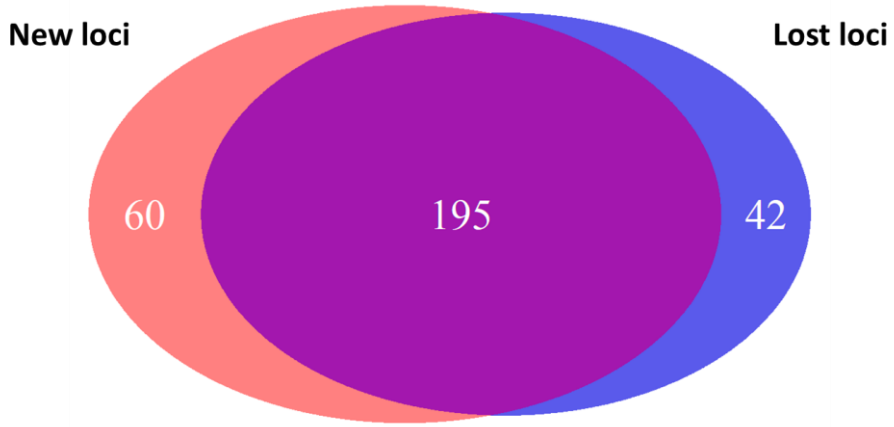


Figure 8. BMI gained and lost loci for the bottom-line analysis in comparison with the minimum p-values reported in the literature

The bottom-line p-values for loci associations with BMI were compared against the minimum-p-values reported. From the 255 significant loci identified by the bottom-line, 60 were newly associated. 42 loci with significant p-values in the original studies, lost their significance with the bottom-line analysis.

Across the 58 traits, 1412 previously published associations appeared to be false positives, after accounting for all studies in the bottom-line analysis. However, 628 associations became significant only after the increase in sample size offered by the bottom-line analysis.

The increment of the sample size achieved by the meta-analysis rises the statistical power, which may lift the significance of some loci while diluting the significance of others. To validate how the sample-size increment affected the significance of the variants; the gained and lost loci were plotted against the effective sample size/max N ratio, which provides a measure of the sample-size increment (Figure 9). Even though there is an outlier, in general, it was observed that the number of gained loci correlates with the sample size increment. This result confirms that by having larger sample sizes, the power of the analysis gets boosted, making possible the discovery of new associations that were previously overseen. Although an equivalent correlation is not observed for the lost loci Vs the sample size increment, it can be seen that for some traits the increment of the sample size may aid in the identification of spurious associations established by smaller studies.

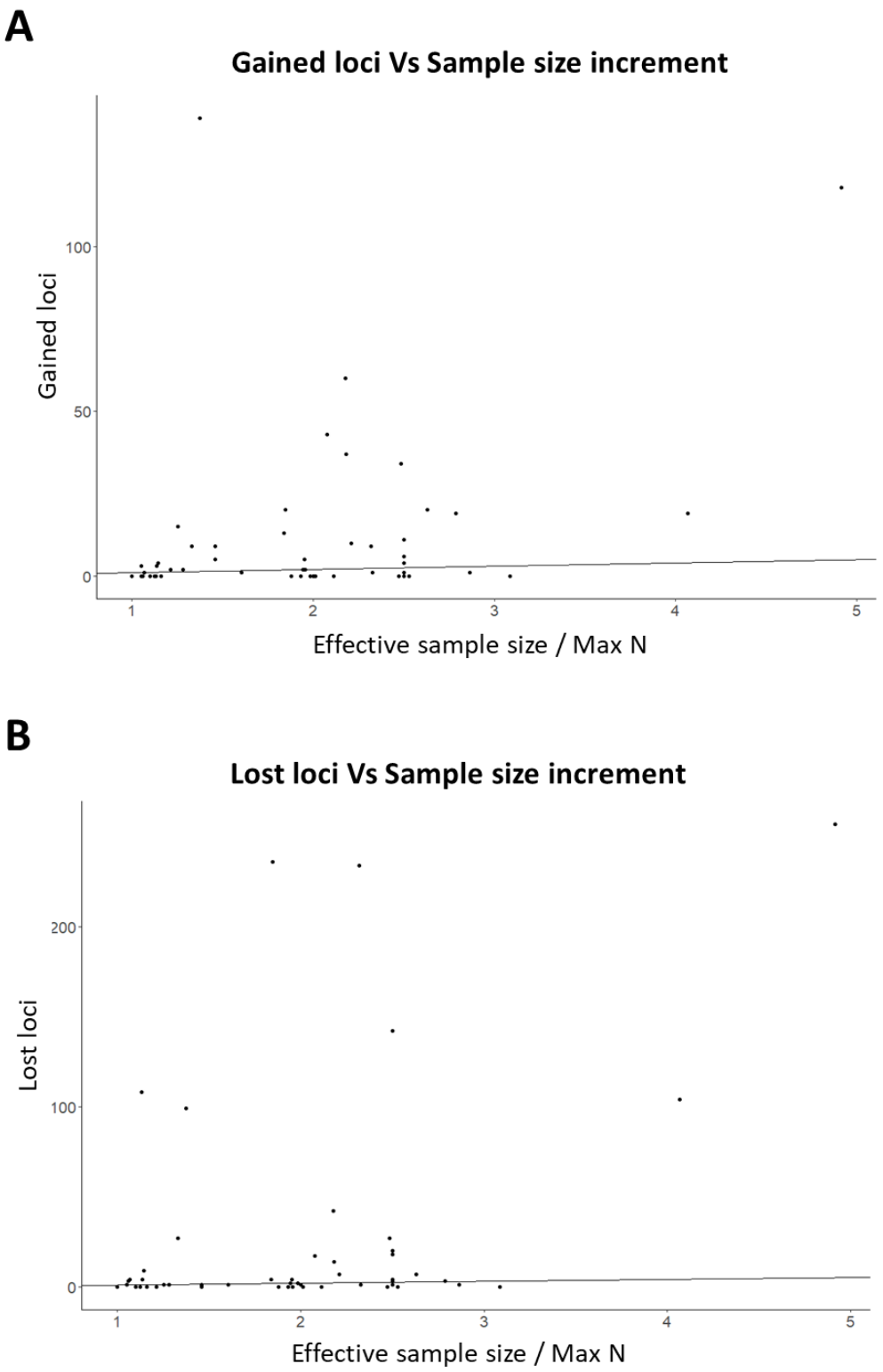


Figure 9. Analysis of the number of gained and lost loci by trait and their correlation to the effective sample size/max N ratio

The gained and lost loci were plotted against the effective sample size/max N ratio, a measurement of the sample size increment achieved by the meta-analysis. A correlation between the sample size increment and the gain loci can be appreciated (A). However, there is no correlation against the lost loci, although, for several traits, even a modest increment of the sample size may aid on the identification of spurious associations.

4.3. ANALYSIS OF THE RELIABILITY OF THE ASSOCIATIONS' STATISTICS PROVIDED BY THE BOTTOM-LINE ANALYSIS

Until now, all analyses included the datasets available in the T2DKP until October 2019. An update of the database was performed in the subsequent months, adding more recently published studies.

In some exceptional cases, some of the old studies which proved to overlap were replaced by a meta-analysis, meaning that the final number of studies could be smaller in May than in October. In general, it was observed that while most of the traits maintained their effective sample size; 7 traits registered a significan increment with an Effective sample size ratio May 2020 / October 2019 > 2 thanks to the newly added datasets, which contained no overlapping subjects. These traits are: BMI, WHR, LDL, SBP, DBP and TG. Stroke_hemorrhagicOn the other hand, 3 exceptional traits reduced their effective sample size, probably due to overlapping studies that were removed from the database (Figure 10).

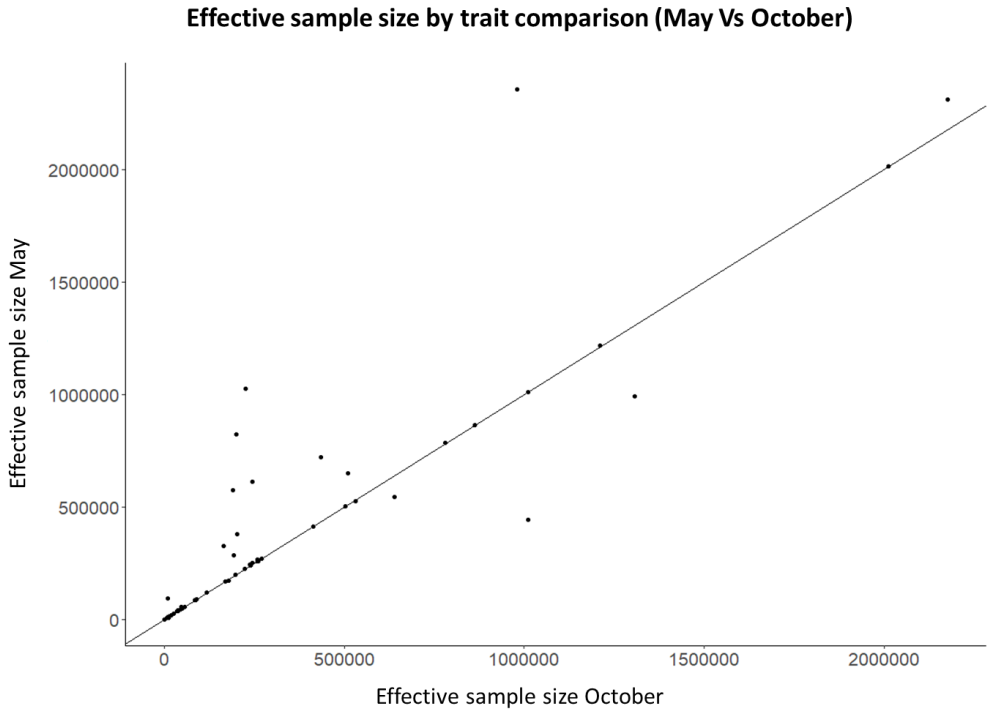


Figure 10. Effective sample size comparison by trait

The figure shows the comparison between the effective sample size computed for the studies available in May in comparison with those that were available in October. It can be observed that while several traits did not modify their effective sample size, some of them increased it, implying that new studies with no overlapping subjects were added to the database. Some exceptional traits reduced their effective sample size, probably due to overlapping studies that were removed.

As a method to validate the reliability of the associations' statistics provided by the bottom-line; the bottom-line p-values were compared against the minimum across studies available to the date of May 2020. Since new studies that increased the sample size were added, it was expected that some of the new associations pointed out by the bottom-line, replicated in the new studies added in May.

When comparing the 60 associations previously discovered for BMI against the minimum p-value of the studies available in May, it was confirmed that 37 (62%) replicated in the most recent BMI GWAS. From 7 traits with a significant increment of their Effective Sample Size (May 2020/October 2019 > 2), 71/104 (68%) loci newly associated by the bottom-line executed in October, replicated in the studies added in May (Figure 11). Contrary to BMI and the other 6 traits with a significant increment of their effective sample size. For T2D (Effective Sample Size ratio May 2020 / October 2019 = 1.1), none of the 118 associated loci by the bottom-line replicated in the studies added in May.

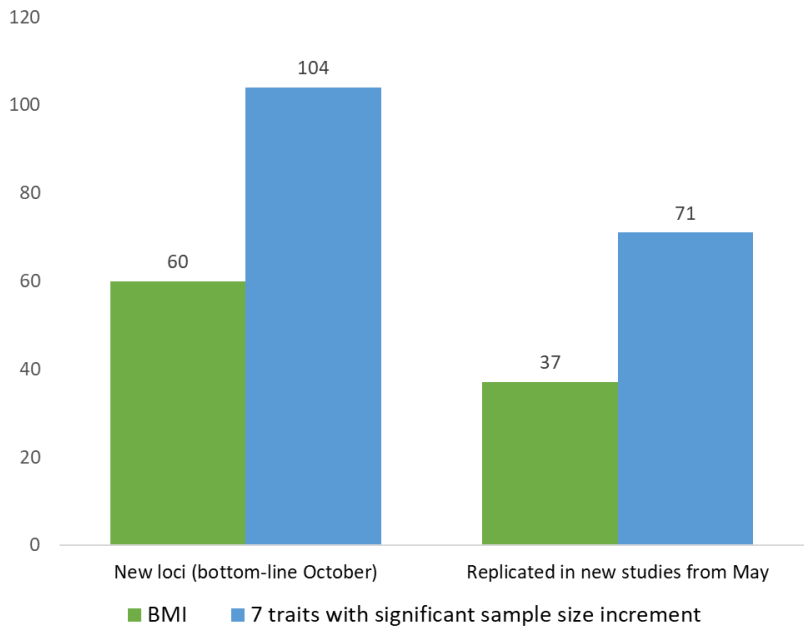


Figure 11. Validation of the loci discovered by the bottom-line analysis
The p-values from the new loci discovered by the bottom-line analysis, that included the studies available until October 2019, were compared against the minimum p-value of the new studies added in May 2020. 37/60 (62%) of the loci associated with BMI replicated in the May studies. From 7 traits with an Effective Sample Size ratio May 2020 / October 2019 >2, 71/104 (68%) loci replicated.

It was previously shown that the naïve meta-analysis provides reasonably well calibrated results when fully overlapping studies are not present. However, the reliability of the loci identified by this approach was tested by comparing their significance against the latter added studies. For

BMI, the naïve meta-analysis identified 22 significant loci that were not significant for the bottom-line analysis. 2 loci, however, were exclusively significant for the naïve meta-analysis in comparison with the bottom-line analysis (Figure 12). 10/22 (45%) replicated in the new studies. The replication rate is lower than the one for the loci discovered by the bottom-line analysis (62%). These results suggest that even when the naïve meta-analysis could be an alternative to the minimum p-value approach, the discovery rate and reliability for the bottom-line analysis are better.

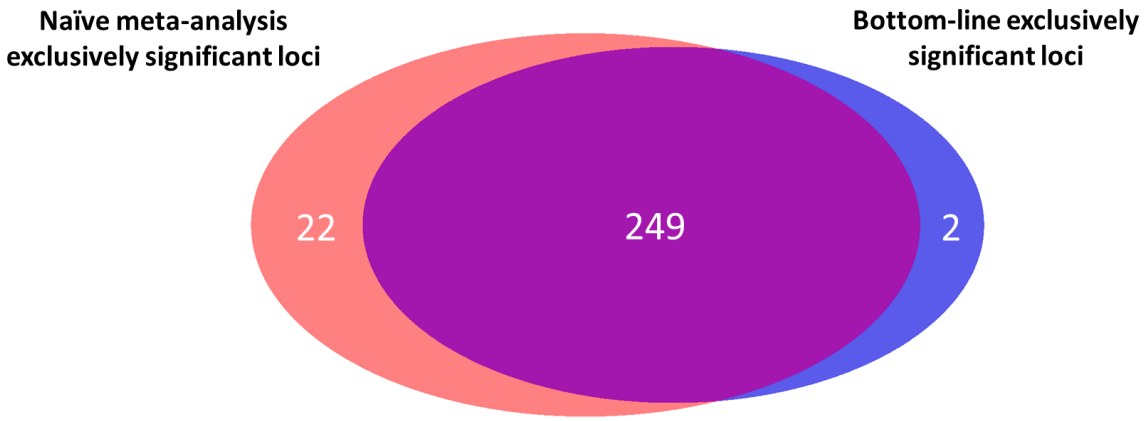


Figure 12. BMI loci significance comparison between the naïve and bottom-line meta-analyses
The bottom-line p-values for loci associations with BMI were compared against the p-values obtained by the naïve meta-analysis. 249 loci were significant for both analyses, 22 loci were exclusively significant for the naïve meta-analysis and 2 loci were exclusively significant for the bottom-line analysis.

4.3.1. PROTEIN-PROTEIN INTERACTION ANALYSES

The 60 loci discovered by the bottom-line were submitted to the DAPPLE software to look for protein-protein interactions among proteins encoded for by genes in the loci. There were 10 disease proteins participating in the direct network (Figure 13A), from which SUFU and FOXO3 have a reported association with BMI in the GWAS Catalog 2019. These results support the associations discovered by the bottom-line, once these genes encoded by the loci have been associated with BMI. The 111 proteins encoded in the 60 loci were submitted to STRING database to perform a protein-protein interaction (PPI) enrichment analysis. The PPI p-value was 0.0128, which means that the network has significantly more interactions than expected (Figure 13B). Such an enrichment indicates that the proteins are at least partially biologically connected, as a group which supports the reliability of the associations discovered by the bottom-line analysis.

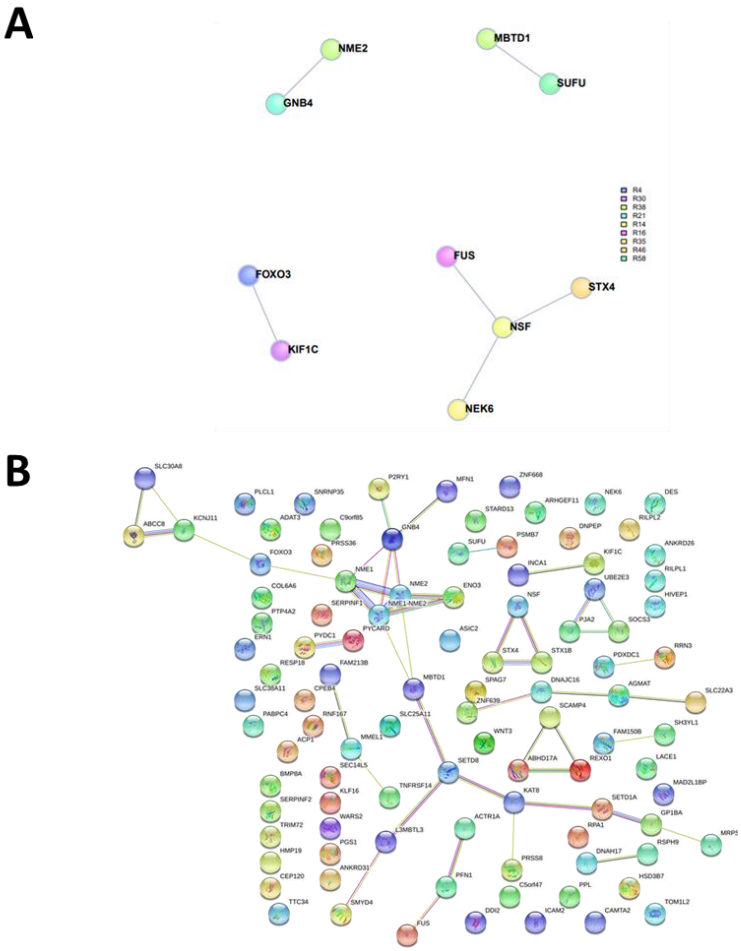
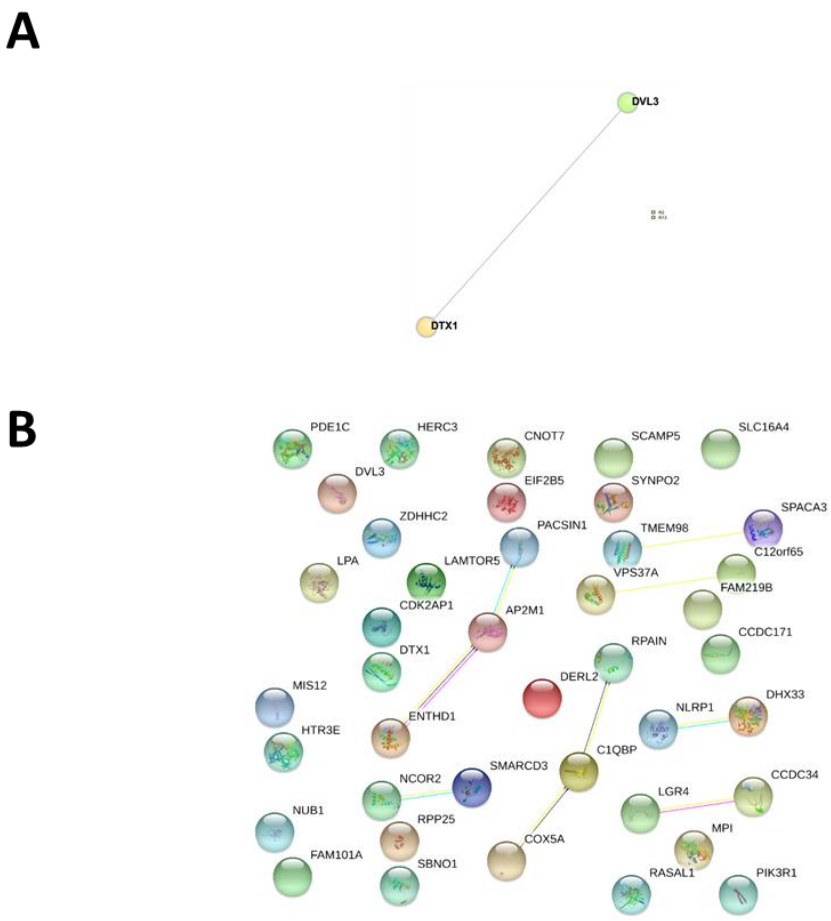


Figure 13. Protein-protein interaction networks of proteins encoded for by genes in the loci associated with BMI by the bottom-line analysis

The 60 loci discovered by the bottom-line analysis were annotated with the software DAPPLE to know the proteins encoded by those regions. 10 disease proteins were identified to participate in a direct network (A). The genes were submitted to STRING database to perform a protein-protein enrichment analysis. The enrichment p-value was 0.0128, which means that the network has more interactions than a random network of the same size, supporting biological meaning of the loci discovered (B).

The protein-protein interaction analyses were perfomed for the 20 genes exclusively discovered by the naïve meta-analysis. 2 disease proteins participated in a direct network, from which DTX1 is reported by the GWAS catalog to be associated with BMI (Figure 14A). The protein-protein enrichment analysis, however, provided an enrichment p-value of 0.159 (Figure 14B), showing that there exist fewer interactions in the network than expected and that the proteins involved have less biological relation.



Network stats

number of nodes:	41	expected number of edges:	6
number of edges:	9	PPI enrichment p-value:	0.159
average node degree:	0.439	*your network does **not** have significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.341		

Figure 14. Protein-protein interaction networks of proteins encoded for by genes in the loci associated with BMI exclusively by the naïve meta-analysis in comparison with the bottom-line analysis

The 22 loci discovered by the bottom-line analysis were annotated with the software DAPPLE to know the proteins encoded by those regions. 2 disease proteins were identified to participate in a direct network (A). The genes were submitted to STRING database to perform a protein-protein enrichment analysis. The enrichment p-value was 0.159, which means that the network has more interactions than a random network of the same size, implying that there is a weak biological relationship among the proteins (B).

Both methods discovered loci encoding genes that have already been reported to be associated with BMI. This outcome validates the reliability of the associations as the analyses were performed without the most recent studies available, for which it is expected that associations for these loci are already reported in the public databases.

The protein-protein interaction results continue to support the better quality of the associations discovered by the bottom-line analysis in comparison with the ones obtained by the naïve meta-analysis. The enrichment p-value obtained for bottom-line proteins is significant while the enrichment analysis obtained by the naïve meta-analysis' proteins is not. These results infer that there are more biological relationships among the proteins encoded by the bottom-line loci than the once expected by chance, increasing the confidence that they share their association with BMI. On the other hand, there is a higher possibility of false-positives within the naïve meta-analysis loci associated, as there are less biological relationships among the proteins encoded by them.

4.4. FALSE POSITIVE AND FALSE NEGATIVE ASSOCIATIONS DISCOVERED WITH THE BOTTOM-LINE ANALYSIS

The bottom-line analysis was performed for a second time, including all the studies available until May 2020 and compared against the minimum p-value to define gained and lost loci. When comparing the number of gained loci against the analysis executed in October 2019, it was observed that in general, the bottom-line from both dates identified a very similar number of associated loci with most of the traits. However, a significant increment in the number of associated loci was produced for the traits for which the sample size was boosted (Figure 15).

Accordingly, the number of lost loci from the October and May analyses were compared. In agreement with what was observed before, the traits with a higher number of lost loci on May in comparison with the studies from October, are those that reported a decrement of the sample size in the newest study. These results confirm the importance of achieving higher sample sizes to be able to discover both new associations and false-positives (Figure 16).

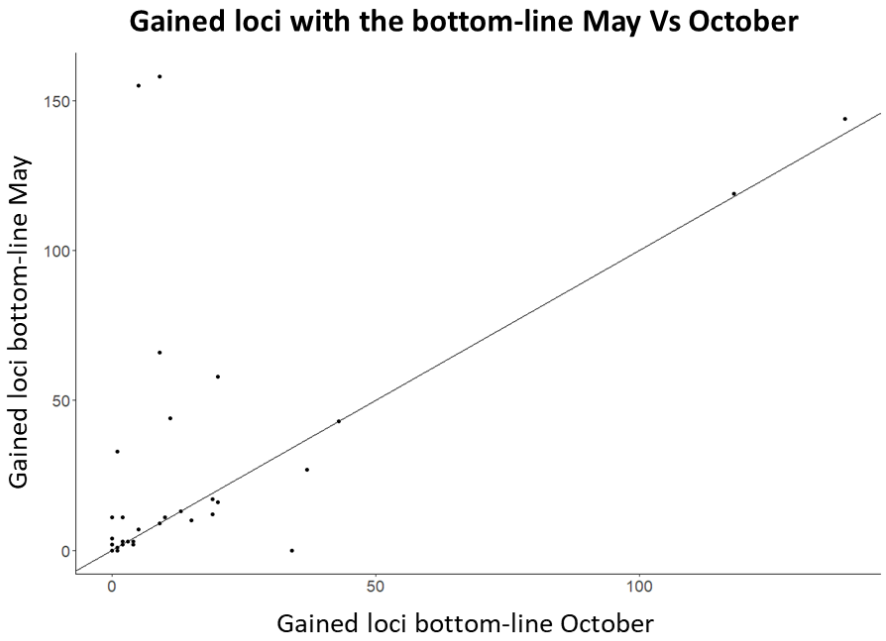


Figure 15. Comparison of the number of newly associated loci with identified from a leave-one-out bottom-line analysis.

The figure shows a comparison of the gained loci when analyzing the studies available in October Vs the analysis with the studies available in May. The significant increment on the number of loci discovered for 7 traits on May in comparison with October, presumably correlates with the traits for which the sample size was boosted.

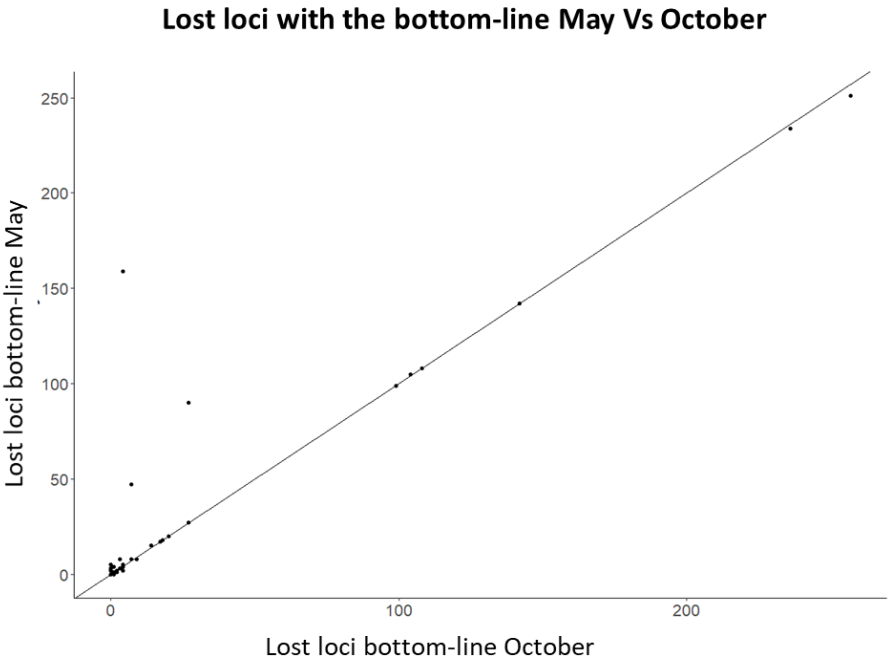


Figure 16. Comparison of the number of lost loci with from a leave-one-out bottom-line analysis.

The figure shows a comparison of the lost loci between the analysis made with the available studies from October Vs the analysis made with the studies available in May. The significant increment on the loci lost for 3 traits on May in comparison with October correlates with the traits for which the sample size was significantly higher when the May studies became included.

Across the 58 traits, 1891 previously published associations appeared to be false positives, after accounting for all studies in the bottom-line analysis. However, 1061 associations were discovered to be significant only after the increase in sample size offered by the bottom-line analysis. A full list of the number of gained and lost loci by trait can be seen in Figure 17, for 3 of the traits, the number of lost loci haven't been computed yet.

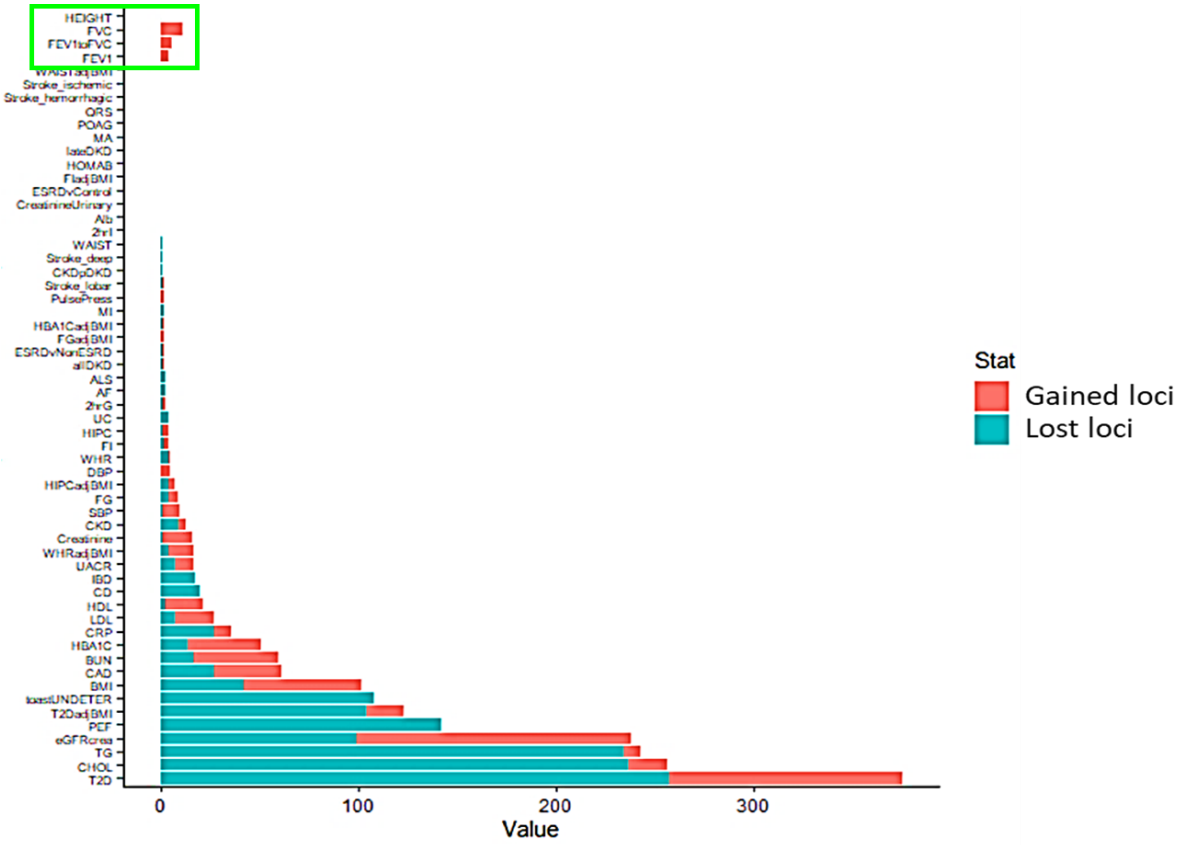


Figure 17. Gained and lost loci by trait

The figure shows the number of new loci associated with each of the traits by the bottom-line analysis. It also shows the number of possible false-positive associations previously reported but lost after the sample size increment offered by the bottom-line analysis. The lost loci for the traits inside the green rectangle haven't been computed yet.

The results suggest that several false-positive associations have been reported in the past. However, a sample overlap aware meta-analysis can provide well-calibrated results, able to discover both false-positive associations and real non previously reported associations.

5. CONCLUSIONS AND FUTURE WORK

The present project validated a well-calibrated overlap-aware meta-analysis strategy, capable of providing uninflated association statistics that take into account all information available. Therefore, increasing the sample size and in consequence, reducing the false-positive and false-negative results. Specifically, it was concluded that:

1. The bottom-line analysis produces non inflated association statistics, even when studies in the analysis fully overlap.
2. A naïve meta-analysis is reasonably well-calibrated when fully overlapping subsets are not included. It could represent an improvement against the minimum p-value strategy to reduce the results' inflation. However, it was observed that the reliability of its associations might not be as high as for the bottom-line.
3. Across the 58 traits analyzed, 1891 previously published associations appeared to be false positives, after accounting for all studies in the bottom-line analysis. However, 1061 associations became significant only after the increase in sample size offered by the bottom-line analysis.
4. This project results suggest that associations reported across published GWAS include a substantial number of false positives and false negatives, but a significant fraction of these can be computationally adjudicated by an overlap-aware meta-analysis strategy.

As future work, it is planned to perform a new analysis in which only the study with the larger sample size is left out as a more standardized method to study the loci discovery replication.

6. BIBLIOGRAPHY

1. Toniolo, A. *et al.* The diabetes pandemic and associated infections. *Rev. Med. Microbiol.* **30**, 1–17 (2019).

2. Kearney, P. M. *et al.* Global burden of hypertension: analysis of worldwide data. *Lancet* **365**, 217–223 (2005).

3. Killin, L. O. J., Starr, J. M., Shiue, I. J. & Russ, T. C. Environmental risk factors for dementia: a systematic review. *BMC Geriatr.* **16**, 1–28 (2016).

4. Wahlsten, D. Genome-wide association studies and the genetic dissection of complex traits. *Am J Hematol* **84**, 159–172 (2019).

5. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33**, 228–237 (2003).

6. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

7. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111**, (2014).

8. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).

9. Altshuler, D. M. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

10. Kim, Y. J. *et al.* A new strategy for enhancing imputation quality of rare variants from next-generation sequencing data via combining SNP and exome chip data. *BMC Genomics* **16**, 1–11 (2015).

11. Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5473–5478 (2015).

12. Barbitoff, Y. A. *et al.* Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Sci. Rep.* **10**, 1–13 (2020).

13. Sobota, R. S. *et al.* Addressing population-specific multiple testing burdens in genetic association studies. *Ann. Hum. Genet.* **79**, 136–147 (2015).

14. Young, A. I. Solving the missing heritability problem. *PLoS Genet.* **15**, 1–7 (2019).

15. Kim, E. E. *et al.* FOLD: A method to optimize power in meta-analysis of genetic association studies with overlapping subjects. *Bioinformatics* **33**, 3947–3954 (2017).

16. Orestis A. Panagiotou, Cristen J. Willer, Joel N. Hirschhorn, and J. P. A. I. The Power of Meta-Analysis in Genome Wide Association Studies. *Bone* **23**, 1–7 (2012).

17. Martorell-Marugan, J., Toro-Dominguez, D., Alarcon-Riquelme, M. E. & Carmona-Saez, P.

MetaGenyo: A web tool for meta-analysis of genetic association studies. *BMC Bioinformatics* **18**, 1–6 (2017).

18. Teri A. Manolio, F. S. C. et. al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

19. Li, A. & Meyre, D. Challenges in reproducibility of genetic association studies: Lessons learned from the obesity field. *Int. J. Obes.* **37**, 559–567 (2013).

20. Type 2 Diabetes Knowledge Portal. Available at: http://www.type2diabetesgenetics.org/. (Accessed: 27th April 2020)

21. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

22. METAL Documentation - Genome Analysis Wiki. Available at: https://genome.sph.umich.edu/wiki/METAL_Documentation#Sample_Overlap_Correction. (Accessed: 27th April 2020)

23. Shah, S. *et al.* Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat. Commun.* **11**, 1–12 (2020).

24. MAHAJAN, A., et al. Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).

25. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).

26. Wuttke M, et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* **51**, 957–972 (2019).

27. Turcot, V. et al. Protein-altering variants associated with body mass index. *Nat Genet* **50**, 26–41 (2018).

28. Teumer, A. *et al.* Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria. *Nat. Commun.* **10**, (2019).

29. Zhao, W, et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat. Genet.* **49**, 1450–1457 (201AD).

30. Horikoshi, M., Kim, Y. J., Korea, S., Moon, S. & Suzuki, K. Identification of type 2 diabetes loci in 433 , 540 East Asian individuals. (2019). doi:10.1101/685172

31. Morris JA, et al. An atlas of genetic influences on osteoporosis in humans and mice. *Nat Genet* **51**, 258–266 (2019).

32. Liu DJ, et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nat Genet* **176**, 139–148 (2017).

33. Justice AE, et al. Protein-coding variants implicate novel genes related to lipid homeostasis contributing to body-fat distribution. *Nat Genet* **51**, 452–469 (2019).

34. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).

35. Ligthart, S. *et al.* Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders. *Am. J. Hum. Genet.* **103**, 691–706 (2018).

36. Yang, J. *et al.* FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490**, 267–273 (2012).

37. Lango Allen H, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–8 (2010).

38. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Mägi R, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).

39. Justice, A. E. *et al.* Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nat. Commun.* **8**, 1–19 (2017).

40. Mahajan A, et al. *Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes*. *Nat Genet* **50**, (2018).

41. Morris, A. P. *et al.* Trans-ethnic kidney function association study reveals putative causal genes and effects on kidney-specific disease aetiologies. *Nat. Commun.* **10**, 1–14 (2019).

42. Graff, M. *et al.* Genome-wide physical activity interactions in adiposity — A meta-analysis of 200,452 adults. *PLoS Genet.* **13**, 1–26 (2017).

43. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).

44. Willer CJ, et al. Discovery and Refinement of Loci Associated with Lipid Levels. *Nat Genet* **45**, 1274–83 (2013).

45. Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–13 (2010).

46. den Hoed M, et al. Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nat Genet* **45**, 621–31 (2013).

47. Scott, R. A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* **66**, 2888–2902 (2017).

48. Scott, R. A., Lagou, V., Welch, R. P., Wheeler, E. & May, E. Europe PMC Funders Group Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. **44**, 991–1005 (2013).

49. Wheeler, E. *et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med.* **14**, 1–30 (2017).

50. Lek M. et al. Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

51. Morris, A. P. *et al.* Large-scale association analysis provides insights into the. *Nat Genet* **44**, 981–990 (2012).

52. Köttgen, A. *et al.* Genome-wide association analyses identify 18 new loci associated with

serum urate concentrations. **45**, 145–154 (2013).

53. Lu, Y. *et al.* New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nat. Commun.* **7**, (2016).

54. Liu, C. T. *et al.* Trans-ethnic Meta-analysis and Functional Annotation Illuminates the Genetic Architecture of Fasting Glucose and Insulin. *Am. J. Hum. Genet.* **99**, 56–75 (2016).

55. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. *The genetic architecture of type 2 diabetes*. *Physiology & behavior* **536**, (2016).

56. Bonàs-Guarch, S. *et al.* Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nat. Commun.* **9**, 1–14 (2018).

57. Chambers, J. C. *et al.* Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat. Genet.* **43**, 1131–1138 (2012).

58. Fritsche, L. G. *et al.* Seven new loci associated with age-related macular degeneration. *Nat. Genet.* **45**, 433–439 (2013).

59. Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).

60. Dastani, Z. *et al.* Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: A multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet.* **8**, (2012).

61. Flannick, J. *et al.* Erratum: Sequence data and association statistics from 12,940 type 2 diabetes cases and controls. *Sci. data* **5**, 180002 (2018).

62. Choi, S. H. *et al.* Monogenic and polygenic contributions to atrial fibrillation risk results from a national biobank. *Circ. Res.* 200–209 (2020). doi:10.1161/CIRCRESAHA.119.315686

63. *et al.* A Genome-Wide Association Study of Diabetic Kidney Disease in Subjects With Type 2 Diabetes. *Diabetes* **67**, 1414–1427 (2018).

64. Sandholm, N. *et al.* The genetic landscape of renal complications in type 1 diabetes. *J. Am. Soc. Nephrol.* **28**, 557–574 (2017).

65. Kilpeläinen, T. O. *et al.* Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels. *Nat. Commun.* **7**, (2016).

66. Khor, C. C. *et al.* Genome-wide association study identifies five new susceptibility loci for primary angle closure glaucoma. *Nat. Genet.* **48**, 556–562 (2016).

67. Ng, M. C. Y. *et al.* Meta-Analysis of Genome-Wide Association Studies in African Americans Provides Insights into the Genetic Architecture of Type 2 Diabetes. *PLoS Genet.* **10**, (2014).

68. Gorski, M. *et al.* 1000 Genomes-based metaanalysis identifies 10 novel loci for kidney function. *Sci. Rep.* **7**, (2017).

69. Sandholm, N., Cole, J. B., Chen, W. & Andrews, D. Genome-wide association study of diabetic kidney disease highlights biology involved in renal basement membrane

collagen. (2018). doi:10.1101/499616

70. Locke, A. E., Steinberg, K. M., Chiang, C. W. K. & Susan, K. Europe PMC Funders Group Exome sequencing of Finnish isolates enhances rare-variant association power. **572**, 323–328 (2020).

71. Chu AY, et al. Multiethnic genome-wide meta-analysis of ectopic fat depots identifies loci associated with adipocyte development and differentiation. *Nat Genet* **49**, 125–130 (2017).

72. Sklar, P. *et al.* Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **43**, 977–985 (2011).

73. Bradfield, J. P. *et al.* A genome-wide association meta-analysis identifies new childhood obesity loci. *Nat. Genet.* **44**, 526–531 (2012).

74. Sim, X. *et al.* Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. *PLoS Genet.* **7**, (2011).

75. Sarnowski, C. *et al.* Impact of Rare and Common Genetic Variants on Diabetes Diagnosis by Hemoglobin A1c in Multi-Ancestry Cohorts: The Trans-Omics for Precision Medicine Program. *Am. J. Hum. Genet.* **105**, 706–718 (2019).

76. SIGMA Type 2 Diabetes Consortium, et al. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97–101 (2014).

77. Laakso, M. *et al.* The Metabolic Syndrome in Men study: A resource for studies of metabolic & cardiovascular diseases. *J. Lipid Res.* **58**, 481–493 (2017).

78. Karpe, F. *et al.* Cohort profile: The Oxford Biobank. *Int. J. Epidemiol.* **47**, 21-21g (2018).

79. Tan, G. D. *et al.* The in vivo effects of the Pro12Ala PPARγ2 polymorphism on adipose tissue NEFA metabolism: The first use of the Oxford Biobank. *Diabetologia* **49**, 158–168 (2006).

80. Hébert, H. L. *et al.* Cohort profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS). *Int. J. Epidemiol.* **47**, 380-381j (2018).

81. Jiang, G. *et al.* Progression of diabetic kidney disease and trajectory of kidney function decline in Chinese patients with Type 2 diabetes. *Kidney Int.* **95**, 178–187 (2019).

82. Wood, A. R. *et al.* A genome-wide association study of IVGTT-based measures of first-phase insulin secretion refines the underlying physiology of type 2 diabetes variants. *Diabetes* **66**, 2296–2309 (2017).

83. Walford, G. A. *et al.* Genome-wide association study of the modified stumvoll insulin sensitivity index identifies BCL2 and FAM19A2 as novel insulin sensitivity loci. *Diabetes* **65**, 3200–3211 (2016).

84. Tan, K. H. X. *et al.* Cohort profile: The Singapore Multi-Ethnic Cohort (MEC) study. *Int. J. Epidemiol.* **47**, 699-699J (2018).

85. Team, A.-D. A. AMP-DCC Quality Control Report. (2019). Available at: https://broad-portal-resources.s3.amazonaws.com/reports/AMP_DCC_QCR_CAMP.v1.0.20190417.0815.pdf.

86. Van Der Heijden, A. A. W. A. *et al.* The Hoorn Diabetes Care System (DCS) cohort. A prospective cohort of persons with type 2 diabetes treated in primary care in the Netherlands. *BMJ Open* **7**, (2017).

87. Dimas, A. S. *et al.* Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes* **63**, 2158–2171 (2014).

88. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

89. Pulit, S. L. *et al.* Meta-Analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).

90. Van Der Harst, P. & Verweij, N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res.* **122**, 433–443 (2018).

91. Cole, J. B., Florez, J. C. & Hirschhorn, J. N. Comprehensive genomic analysis of dietary habits in UK Biobank identifies hundreds of genetic associations. *Nat. Commun.* **11**, 1–11 (2020).

92. Gazal, S. *et al.* Functional Architecture of Low-Frequency Variants Highlights Strength of Negative Selection Across Coding and Non-Coding Annotations. **50**, 1600–1607 (2019).

93. Forgetta V, Jiang L, Vulpescu NA, et al. An Effector Index to Predict Causal Genes at GWAS Loci.

94. Ntalla, I. *et al.* Multi-ancestry GWAS of the electrocardiographic PR interval identifies 202 loci underlying cardiac conduction. *Nat. Commun.* **11**, 1–12 (2020).

95. Surendran P, et al. Trans-ancestry Meta-Analyses Identify Rare and Common Variants Associated With Blood Pressure and Hypertension. *Nat Genet* **48**, 1151–1161 (2016).

96. FinnGen. FinnGen Data Freeze 2. (2020). Available at: https://www.finngen.fi/en/access_results.

97. Pirruccello, J. P. *et al.* Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. *Nat. Commun.* **11**, 1–10 (2020).

98. Liu, C. T. *et al.* Genome-wide Association Study of Change in Fasting Glucose over time in 13,807 non-diabetic European Ancestry Individuals. *Sci. Rep.* **9**, 1–8 (2019).

99. Gaulton, K. J. *et al.* Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).

100. Pollack, S. *et al.* Multiethnic genome-wide association study of diabetic retinopathy using liability threshold modeling of duration of diabetes and glycemic control. *Diabetes* **68**, 441–456 (2019).

101. Anthony L. Hinrichs, E. K. L. and B. K. S. Population Stratification and Patterns of Linkage Disequilibrium Anthony. **33**, S88–S92 (2009).

7. APPENDIX

7.1. METAL MASTER SCRIPT

```
#Parameters
SEPARATOR TAB
COLUMNCOUNTING LENIENT
#AVERAGEFREQ ON #Uncomment to track allele frequencies
#MINMAXFREQ ON #Uncomment to track allele frequencies

#Identification of the files
MARKER   varId
ALLELE   reference alt
WEIGHT   n
EFFECT   beta
STDERR   stdErr
PVAL     pValue
CUSTOMVARIABLE TotalSampleSize
LABEL TotalSampleSize AS n
#FREQLABEL maf #Uncomment to track allele frequencies
```

7.2. BOTTOM-LINE META-ANALYSIS IMPLEMENTATION CODE

```
# !/bin/bash
#Run example ./Programs/BottomLineImp.sh BMI ./Data Oct BOct_BMI.txt
#Requires ./Programs/MasterFileMetal.txt and



#Checking that the number of parameters passed is correct
if [ "$#" -ne 4 ]; then
    echo $# "Not the right number of parameters"
    echo $1
    echo $2
    exit 2
    fi


#Main variables
trait=$1
MainDir=$2
Type=$3
CurrentList=$4 #List of studies to which the bottom-line results will
be compared to
MmasterFile=./Programs/MasterFileMetal.txt
```

```
MmasterFileMAF=./Programs/MasterFileMetalMAF.txt
FinalMasterFile=./Programs/MasterFileMetalBottomline.txt
ParentalDir=/humgen/diabetes2/users/mvg/portal/scripts/VARIANTS/meta_d
ata_file.ver4-META_MDV.txt

#Uncomment next line if customized Metal Master file required
#MmasterFile=$6

mkdir -p ./Results/Metal/$1
DirForMetal=./Results/Metal/$1
MPath=$DirForMetal/$Type

#Dir to store the Big final fies
mkdir -p ./Results/RFilesToUse/$1
EndDir=./Results/RFilesToUse/$1


#Store the path to the Files Directory
FilesDir=$MainDir/$1/Files/FilesToUse

#Save all the studies file names
studiesNames=$( ls $FilesDir | grep ".tsv" )

#Save the name of all the Bottomline studies
Botstudies=$( cat $MainDir/$1/$CurrentList)



#******Bottomline Analysis Implementation **************



###**********************Common and Rare Split******************

#Make a dir to save the MAF split files
FilesDirMAF=$FilesDir/MAF
mkdir -p $FilesDirMAF

#Save the names of the paths for the Common and the rare variant files
studiesPathsC=()
studiesPathsR=()
studiesPaths=()

actualAns=()


for s in $Botstudies;do
    studiesPathsC+=($FilesDirMAF/$1_${s}_Common.tsv)
    studiesPathsR+=($FilesDirMAF/$1_${s}_Rare.tsv)
    studiesPaths+=($FilesDir/$1_${s}.tsv)
    sPath=($FilesDir/$1_${s}.tsv)



    #Find automatically the column for MAF
    colMaf=$(cat $sPath | awk -F '\t' -v col='maf' 'NR==1{for (i=1;
i<=NF; i++) if ($i==col) {print i;exit}}')
    colP=$(cat $sPath | awk -F '\t' -v col='pValue' 'NR==1{for (i=1;
i<=NF; i++) if ($i==col) {print i;exit}}')
    colB=$(cat $sPath | awk -F '\t' -v col='beta' 'NR==1{for (i=1; i<=NF;
i++) if ($i==col) {print i;exit}}')
```

```
    #If the file hasn't been split yet, make the split
    studyC=$FilesDirMAF/$1_${s}_Common.tsv
    if [ ! -f $studyC ]; then
        echo $studyC "File not found!"
        cat $sPath | awk -F "\t" '{if($14 >= 0.05 || $14 == ""){ print
} } ' OFS="\t" | awk -F'\t' '$10 !="" ' OFS='\t' | awk -F'\t' '$11 !=""
' OFS='\t' > $studyC
    fi

    studyR=$FilesDirMAF/$1_${s}_Rare.tsv
    if [ ! -f $studyR ]; then
        echo $studyR "File not found!"
        cat $sPath |  awk -F '\t' '{if($14 < 0.05 || $14 == "maf"){print
}}' OFS='\t'| awk -F '\t' '$14 != ""' OFS='\t' | awk -F'\t' '$10 !=""'
OFS='\t' | awk -F'\t' ' $11!="" ' OFS='\t'  > $studyR
    fi


    #Save the file to unite
    studyRJ=$FilesDirMAF/$1_${s}_Rare_join.tsv
    tail -n +2 $studyR | cut -f1,2,3,4,5,10,11,15,17 > $studyRJ


    #Find all the repeated ancestries
    colAncestry=$(cat $sPath | awk -F '\t' -v col='ancestry' 'NR==1{for
(i=1; i<=NF; i++) if ($i==col) {print i;exit}}')

    Ans=$(head -2 $sPath | cut -f$colAncestry | tail -n1)

    actualAns+=( $Ans )

done


#Find the unique ancestries

UniqAncestries=($(echo "${actualAns[@]}" | tr ' ' '\n' | sort -u | tr
'\n' ' '))


#Save the final ancestries (If more than 1 ancestry present, remove
mixed. If only mixed present, the AncestryArray will be empty, and the
only analysis will be with overlap off

FinalAncestries=()

if [ ${#UniqAncestries[@]} -gt 1 ];then

    for Ancestry in ${UniqAncestries[@]};do
        if [ "$Ancestry" != "Mixed" ];then
            FinalAncestries+=( $Ancestry )
        fi
    done

    if [ ${#FinalAncestries[@]} -gt 0 ];then

        AncestryArray=( "${FinalAncestries[@]}" )

    fi
```

```
else
    AncestryArray=( "${UniqAncestries[@]}" )
fi

echo $trait

echo "Original ancestries: " ${UniqAncestries[*]}
echo "Final ancestires: " ${AncestryArray[*]}




###Separate the files by Minor Allele Frequency

for Ancestry in ${AncestryArray[@]};do

    #For each study:

    #Get a list of the ancestry specific files
    echo $Ancestry

    #**** AncestryFiles Common will save all the ancestry related files
    AncestryFilesC=()
    for study in ${studiesPathsC[@]};do

        colAncestry=$(cat  $study  |  awk  -F  '\t'  -v  col='ancestry'
'NR==1{for (i=1; i<=NF; i++) if ($i==col) {print i;exit}}')

        Ans=$(head -2 $study | cut -f$colAncestry | tail -n1)

        if [[ $Ans == $Ancestry ]];then
            AncestryFilesC+=( $study )

        fi

    done



    ##### Get Metal file and Run Metal for the COMMON VARIANTS OF THE
ANCESTRY SAMPLESIZE
    mkdir -p $DirForMetal/$Type/

    cp $MmasterFile $DirForMetal/$Type/Run_$1_$Ancestry-Common-Samp.txt
    MFilePathCS=$DirForMetal/$Type/Run_$1_$Ancestry-Common-Samp.txt

    MPath=$DirForMetal/$Type

    echo -e "\nOUTFILE" $1_$Ancestry-Common-Samp .tbl >> $MFilePathCS
    echo -e "\nSCHEME SAMPLESIZE\n" >> $MFilePathCS
    echo -e "\nOVERLAP ON\n" >> $MFilePathCS

    for file in "${AncestryFilesC[@]}";do
        echo "PROCESS" $file >> $MFilePathCS
    done

    echo -e "\nANALYZE\nQUIT" >> $MFilePathCS

    #Run Metal
    echo -e "Running Metal for" $1_$Ancestry-Common-Samp
```

```
    ########    change path for metal   ####################
    ./Programs/metal $MFilePathCS

    mv     $1_$Ancestry-Common-Samp1.tbl     $MPath/$1_$Ancestry-Common-
Samp.tbl
    mv  $1_$Ancestry-Common-Samp1.tbl.info  $MPath/$1_$Ancestry-Common-
Samp.tbl.info



    ##### Get Metal file and Run Metal for the COMMON VARIANTS OF THE
ANCESTRY STDERR
    mkdir -p $DirForMetal/$Type/

    cp $MmasterFile $DirForMetal/$Type/Run_$1_$Ancestry-Common-Err.txt
    MFilePathCE=$DirForMetal/$Type/Run_$1_$Ancestry-Common-Err.txt

    MPath=$DirForMetal/$Type

    echo -e "\nOUTFILE" $1_$Ancestry-Common-Err .tbl >> $MFilePathCE
    echo -e "\nSCHEME STDERR\n" >> $MFilePathCE
    echo -e "\nOVERLAP OFF\n" >> $MFilePathCE

    for file in "${AncestryFilesC[@]}";do
        echo "PROCESS" $file >> $MFilePathCE
    done

    echo -e "\nANALYZE\nQUIT" >> $MFilePathCE

    #Run Metal
    echo -e "Runing Metal for" $1_$Ancestry-Common-Err
    ./Programs/metal $MFilePathCE

    mv $1_$Ancestry-Common-Err1.tbl $MPath/$1_$Ancestry-Common-Err.tbl
    mv   $1_$Ancestry-Common-Err1.tbl.info   $MPath/$1_$Ancestry-Common-
Err.tbl.info




    #Unite the two Analyses (SampleSize and Error)

    join -1 1 -2 1 -o 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.4 2.5 <(sort
$MPath/$1_$Ancestry-Common-Samp.tbl    |    tail    -n    +2)    <(sort
$MPath/$1_$Ancestry-Common-Err.tbl | tail -n +2 ) > $MPath/$1_$Ancestry-
Common.tmp

    #sometimes METAL will flip the alleles. Check if that is the case,
and if necessary, flip them back together with the z-score
    cat $MPath/$1_$Ancestry-Common.tmp | tr -s ' ' | tr ' ' '\t' |awk -
F'\t'  '$1!="MarkerName"{print}'  OFS="\t"|  awk  -F'\t'  '{print  $1,$0}'
OFS='\t'      |           awk      -F'\t'      '{      gsub(/:/,"\t",$2);
sub(/[[:lower:]]/,toupper($3),$3);   sub(/[[:lower:]]/,toupper($4),$4);
print}'     OFS='\t'     |     awk     -F'\t'     '{if($4==$6)     {print
$1,$2,$3,$4,$5,$11,$14,$15,$13}          else          if($4!=$6){print
$1,$2,$3,$4,$5,$11,$14*-1,$15,$13}}'  OFS='\t'  >  $MPath/$1_$Ancestry-
Common.txt
```

```
    sed                                -i                                '1i
varId\tchromosome\tposition\treference\talt\tpValue\tbeta\tstdErr\tn'
$MPath/$1_$Ancestry-Common.txt

    rm $MPath/$1_$Ancestry-Common.tmp


    ######## Concatenate the Rare variants from the ancestry

    #A final file with all the rare united variants will be stored in
this file
    rm $MPath/$1_All_Rare_$Ancestry.tsv
    touch $MPath/$1_All_Rare_$Ancestry.tsv


    for s in $Botstudies;do

        colAncestry=$(cat $FilesDirMAF/$1_${s}_Rare.tsv | awk -F '\t' -
v  col='ancestry'  'NR==1{for  (i=1;  i<=NF;  i++)  if  ($i==col)  {print
i;exit}}')

        Ans=$(head -2 $FilesDirMAF/$1_${s}_Rare.tsv | cut -f$colAncestry
| tail -n1)

        if [[ $Ans == $Ancestry ]];then
            cat          $FilesDirMAF/$1_${s}_Rare_join.tsv          >>
$MPath/$1_All_Rare_$Ancestry.tsv
        fi

    done



###################Unite the common and rare variants together and keep
the one with the biggest n


    rm $MPath/$1_$Ancestry.txt

    cat $MPath/$1_$Ancestry-Common.txt $MPath/$1_All_Rare_$Ancestry.tsv
|   sort -k1,1 -k9,9gr | awk -F'\t' '!a[$1] {a[$1] = $9} $9 == a[$1]'
OFS='\t'           |           awk           -F'\t'           '{printf
"%s\t%s\t%s\t%s\t%s\t%.1e\t%1.4f\t%1.4f\t%d\n",$1,$2,$3,$4,$5,$6,$7,$8
,$9   }' | uniq | awk -F'\t' '$1 !="varId" && $6 > 0' OFS='\t' >
$MPath/$1_$Ancestry.txt
    sed                                          -i                '1i
varId\tchromosome\tposition\treference\talt\tpValue\tbeta\tstdErr\tn'
$MPath/$1_$Ancestry.txt


done



###### MAKE THE FINAL METAL WITH OVERLAP OFF SAMPLE SIZE ########


cp $MmasterFile $DirForMetal/$Type/Run_$1_$Type-Samp.txt
MFilePathFS=$DirForMetal/$Type/Run_$1_$Type-Samp.txt
```

```
echo -e "\nOUTFILE" $1_$Type-Samp .tbl >> $MFilePathFS
echo -e "\nSCHEME SAMPLESIZE\n" >> $MFilePathFS
echo -e "\nOVERLAP OFF\n" >> $MFilePathFS

for Ancestry in ${AncestryArray[@]};do
    CFile=$DirForMetal/$Type/$1_$Ancestry.txt
    if [ -f $CFile ];then
        echo "PROCESS" $CFile >> $MFilePathFS
    fi

done

echo -e "\nANALYZE\nQUIT" >> $MFilePathFS

#Run Metal
echo -e "Runing Metal for" $1_$Type
./Programs/metal $MFilePathFS

mv $1_${Type}-Samp1.tbl $MPath/$1_$Type-Samp.tbl
mv $1_${Type}-Samp1.tbl.info $MPath/$1_$Type-Samp.tbl.info



######## MAKE THE FINAL METAL WITH OVERLAP OFF ERR ###########


cp $MmasterFile $DirForMetal/$Type/Run_$1_$Type-Err.txt
MFilePathFE=$DirForMetal/$Type/Run_$1_$Type-Err.txt


echo -e "\nOUTFILE" $1_$Type-Err .tbl >> $MFilePathFE
echo -e "\nSCHEME STDERR\n" >> $MFilePathFE
echo -e "\nOVERLAP OFF\n" >> $MFilePathFE

for Ancestry in ${AncestryArray[@]};do
    CFile=$DirForMetal/$Type/$1_$Ancestry.txt
    if [ -f $CFile ];then
        echo "PROCESS" $CFile >> $MFilePathFE
    fi

done

echo -e "\nANALYZE\nQUIT" >> $MFilePathFE

#Run Metal
echo -e "Runing Metal for" $1_$Type
./Programs/metal $MFilePathFE

mv $1_${Type}-Err1.tbl $MPath/$1_$Type-Err.tbl
mv $1_${Type}-Err1.tbl.info $MPath/$1_$Type-Err.tbl.info



#Unite the two Analyses (SampleSize and Error)

join -1 1 -2 1 -o 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 2.4 2.5 <(sort
$MPath/$1_$Type-Samp.tbl)   <(sort   $MPath/$1_$Type-Err.tbl      )   >
$MPath/$1_$Type.tmp
```

```
#sometimes METAL will flip the alleles. Check if that is the case, and
if necessary, flip them back together with the z-score
cat  $MPath/$1_$Type.tmp  |  tr  -s  '  '  |  tr  '  '  '\t'  |awk  -F'\t'
'$1!="MarkerName"{print}' OFS="\t"| awk -F'\t' '{print $1,$0}' OFS='\t'
|   awk -F'\t' '{ gsub(/:/,"\t",$2); sub(/[[:lower:]]/,toupper($3),$3);
sub(/[[:lower:]]/,toupper($4),$4);   print}'   OFS='\t'   |   awk   -F'\t'
'{if($4==$6)   {print   $1,$2,$3,$4,$5,"BMI",$10,$13,$9,$14,$12}   else
if($4!=$6){print       $1,$2,$3,$4,$5,"BMI",$10,$13*-1,$9*-1,$14,$12}}'
OFS='\t'|                awk                -F'\t'                '{printf
"%s\t%s\t%s\t%s\t%s\t%s\t%.1e\t%1.4f\t%1.4f\t%1.4f\t%d\n",$1,$2,$3,$4,
$5,$6,$7,$8,$9,$10,$11  }' > $MPath/$1_$Type.tbl


sed                                -i                                  '1i
varId\tchromosome\tposition\treference\talt\tphenotype\tpValue\tbeta\t
zScore\tstdErr\tn' $MPath/$1_$Type.tbl

rm $MPath/$1_$Type.tmp



#########  MAKE  METAL  WITH  OVERLAP  OFF  WITH  THE  ORIGINAL  FILES
############

#cp $MmasterFile $DirForMetal/$Type/Run_$1_${Type}_Naïve.txt
cp $MmasterFileMAF $DirForMetal/$Type/Run_$1_${Type}_Naïve.txt
MFilePathO=$DirForMetal/$Type/Run_$1_${Type}_Naïve.txt


echo -e "\nOUTFILE" $1_${Type}_Naïve .tbl >> $MFilePathO
echo -e "\nOVERLAP OFF\n" >> $MFilePathO

for FFile in ${studiesPaths[@]};do
    echo "PROCESS" $FFile >> $MFilePathO
done


echo -e "\nANALYZE\nQUIT" >> $MFilePathO

#Run Metal
echo -e "Runing Metal for" $1_${Type}_Naïve
./Programs/metal $MFilePathO

mv $1_${Type}_Naïve1.tbl $MPath/$1_${Type}_Naïve.tbl
mv $1_${Type}_Naïve1.tbl.info $MPath/$1_${Type}_Naïve.tbl.info
```